Towards Health Management Intelligence:

A Case Study from South Africa

A DISSERTATION PRESENTED TO THE DEPARTMENT OF INFORMATION SYSTEMS

AT THE UNIVERSITY OF CAPE TOWN

By

GREGORY ROWLES (RWLGRE001)

In partial fulfilment of the requirements for Master of Commerce in Information Systems (INF5005W)

Declaration

I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

I have used the APA convention for citation and referencing.

Each contribution to, and quotation in, this paper from the works of other people has been attributed, and has been cited and referenced.

This "Towards Health Management Intelligence: A Case Study from South Africa" is my own work

I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work

Signature: Signed by candidate
Signature Removed

Date: 30/11/2014

Full Name of Student: Gregory Thomas Rowles

Student Number: RWLGRE001

ABSTRACT

Over the last two decades various information management processes have evolved in South Africa's public health system. Most notably a self-service business intelligence tool has emerged at the national level which has been supported by the presence of a Routine Health Information System. Corporate business intelligence and its underlying process are well documented but not in the public health domain. The emergence of this tool and the underlying support processes are investigated in a longitudinal case study. Complex adaptive systems theory is used to demonstrate the evolutionary path of business intelligence processes according to four key areas, namely data quality, master data management, data warehousing and analytics. These processes have developed out of an information management culture that has been nurtured by a participatory approach which required an attractor: the improvement of health services through the collection and use of information. The evolution of these processes took place through a bottom up approach that relied on distributed control structures, self-organization and regular engagement within the CAS that is South Africa's public health system. This created an environment in which information quality practices and master data management processes enabled the continued production of data for warehousing and analytics. Findings will show how business intelligence processes have evolved within a public health setting to the point that they are supported by a new policy that ensures data integrity, presence, quality and use processes. These processes have developed and stabilized over many iterations and have enabled the establishment of a country level self-service business intelligence platform for health managers.

ACKNOWLEDGEMENTS

To my loving wife Marica for her support and motivation that carried me through this process. To my family for their loving support and encouragement. To my colleagues at the Health Information Systems Programme and at the Department of Health; without your insights and support this research project would not be possible. To those within the HISP international community who have inspired me, Skål. To Calle Hedberg for being a mentor, role model, and guide. To Vincent Shaw for his leadership, support and guidance. To Francois Venter whose insights and decision to develop a self-service BI tool would have far reaching consequences. To Prof Jean-Paul van Belle for making this journey possible and especially to Prof Mike Hart for his patience, dedication and perseverance in thought.

Thank you all for making this experience possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AG	Auditor General
APP	Annual Performance Planning
BI	Business Intelligence
BI&A	Business Intelligence and Analytics
BIS	Business Intelligence Systems
BU	Business Users
CAS	Complex Adaptive Systems
CDC	Centre for Disease Control
CI	Competitive Intelligence
CIO	Chief Information Officer
DBMS	Database Management System
DHIS	District Health Information System
DHMIS	District Health Management Information System
DIO	District Information Officer
DOH	Department of Health
DQ	Data Quality
DQM	Data Quality Management
DW	Data Warehousing
ETL	Extract, Transform and Load
EHR	Electronic Health Record
EMR	Electronic Medical Record
HIS	Health Information Systems
HMIS	Health Management Information System
GIS	Geographic Information System
ICT	Information and Communication Technology
IQ	Information Quality
MDM	Master Data Management
OLAP	Online Analytical Processing
NHIRD	National Health Information Repository and Data Warehouse
NHISSA	National Health Information System for South Africa
NIDS	National Indicator Data Set
NDOH	National Department of Health
PHC	Primary Health Care
PIDS	Provincial Indicator Data Set
PIO	Provincial Information Officer
QRS	Quarterly Reporting System

RDMS	Relational Database Management System
RHIS	Routine Health Information Systems
SASQAF	South African Statistical Quality Assurance Framework
SSBI	Self-Service Business Intelligence
TQM	Total Quality Management
TS	Time Series
UCT	University of Cape Town
USAID	United States Agency for International Development
UWC	University of Western Cape

1. Introduction

Various projects, programmes and initiatives are underway in South Africa's public health system, all serving different administrative and health management concerns. These efforts result in the collection and storage of health service data, much of which is used to assess health outcomes. Since 1994 South Africa's Department of Health (DOH), together with international donors and non-profit organizations, have supported the development of information systems (IS) to provide this data. By integrating and analysing health information managers, practitioners, development partners and a wide range of health professionals have also been able to plan and manage the provision of health services.

The recent decision by the National Department of Health (NDOH) to implement a project unit that seeks to consolidate and present health information in a national data warehouse raises many questions about the design of health system information processes. The 'NHIRD' project unit was initiated by NDOH with the goal of acting as a centralized repository for health information in order to provide 'intelligence' to health system decision makers. It includes the implementation of a self-service reporting tool designed and adapted locally to make health data available to managers. While the development and endorsement of this project unit by NDOH could be considered a success indicator for its underlying information systems, it raises many questions about the information processes contributing to this project unit. There appears to be some consistency with BI processes typically found in corporate enterprises.

Most organizations invest in business intelligence (BI) in order to support and improve decision making. This typically requires the functioning of a diverse set of processes to ensure data is present, relevant, of high quality, integrated and capable of delivering insight when required. These processes are prevalent in corporate environments and are a top priority for executive level managers. Country health systems are very different, large in scale and have diverse challenges that are managed with constrained budgets. While the establishment and endorsement of this project unit by NDOH could be considered a success indicator for its underlying information systems, it raises many questions about information gathering and management processes in resource constrained environments as large as the South African DOH. The leap into country level data warehousing could not take place overnight and had to rely on the presence of well-functioning mechanisms and processes.

The question remains whether or not corporate BI processes are generic, relevant and present in resource constrained settings such as a public health system. Are they present, how have they evolved and what lessons are there for similar implementations in other countries?

The research objectives are clear: demonstrate how BI processes have evolved in a complex adaptive system; use CAS theory to untangle the BI processes and demonstrate that business centric phenomena of a corporate nature have taken place, are prevalent and exist within other IS domains that have advanced above a minimum level of information management processing.

The following chapters look at BI from various perspectives and discuss their four major focus areas in detail. These include information quality, master data management, data warehousing and analytics. Complex adaptive systems (CAS) theory has been selected as the theory for interpreting mechanisms of BI. The review into CAS theory will be followed by a brief analysis of Health Information Systems (HIS) as the longitudinal case study finds itself evolving from within this IS domain. Discussions will begin with a personal account of the researcher who participated in the development of the underlying information system. This is followed by a review of the emerging BI focus areas that took shape over the years. An attempt will be made to untangle and demonstrate how these mechanisms of BI evolved due to a complex adaptive design enabled by agents of the system. Key learnings will be brought forward. The final chapter is a discussion that concludes with findings, limitations, recommendations and areas for further research.

2. Literature Review

The following literature review is an exploration of business intelligence areas that are typically associated with industry but will be focussed on BI from the health systems perspective which is more relevant to the case study. This section starts off with a review into the function of intelligence, consciousness and our need for integrated data. This is followed up by a brief review of BI objectives, architectures, analysis processes and social models that currently apply to the area. Competitive intelligence is included in this review to demonstrate its similarities. Major attention has been paid to information quality, master data management, data warehousing and analytics because events unfolding throughout the development of the system showed how crucial they were to the health context which helped to achieve suitable outputs and analytics. This is followed up by a brief review of complex adaptive systems theory and is rounded off with a review into health information systems. This is done to provide background to the environment in which the emerging system is found.

2.1 Business Intelligence

With persistent drops in technology costs over the years it has become possible for more and more data to be stored and aligned in ways that were previously unfeasible. Cohen (1999) suggested we were heading towards a complexity revolution that would provide new and unique opportunities to utilize complex ties of inter-process data. Information intelligence may be the realization of that vision as it provides us with a means to focus our technology and efforts on the activities related to the collection, integration, analysis and use of information to assess our advancements and measure the achievement of our goals.

2.1.1 Intelligence, Consciousness and the need for Integrated Data

In human psychology "[Intelligence] . . . involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather it reflects a broader and deeper capability for comprehending our surroundings—'catching on', 'making sense' of things, or 'figuring out' what to do" (Gottfredson, 1997, p. 13). Waser (2011, p. 2) provides an evidence based description for intelligence as "a measure of the ability to determine how to achieve a wide variety of goals under a wide variety of circumstances". Waser (2011) also describes the *function* of intelligence as the ability to determine methods by which various goals can be achieved under a wide range of circumstances. It is measured by the level of information processing required to manipulate circumstances so that they include the goal. The lower the information processing required to influence outcomes (so that they include the goal), the higher the degree of intelligence.

Tononi (2004) introduced the *information-integration-theory-of-consciousness* that associates consciousness directly with information structures and information processing. The theory suggests that different levels of consciousness exist with different levels of experience and the quality of the experience is proportionate to the level of consciousness. These all relate to information-producing structures and information-integrating capabilities. The higher the amount of information generated by complex elements coupled with a high degree of informational-relationship capabilities, the higher the quality of experience (Waser, 2011). The ability of a system to integrate its information will grow as that system incorporates statistical-regularities within its environment, thereby creating opportunities for learning (Tononi, 2004; Waser, 2011).

It is worth recognizing that complex adaptive systems (CAS) theory focuses primarily on people as the active components within systems, and these agents and coalitions are the ultimate decision makers that influence, guide and support our organizations. The ability for systems to achieve elevated cognitive-states of awareness, and their ability to respond appropriately has significant bearing on our organizations and enterprises. Without effective information generation, integration and dissemination services our ability to understand and guide our organizations through changing circumstances is impeded.

One can conclude that our organizations require effective information design, generation, exchange, integration and internalization mechanisms across various processes and between various agents to support and elevate levels of awareness. When combined with Boisot and Child's (1999) views on adaptation in complex environments, accessible and 'integrable' information should lead to organizational awareness or consciousness, which allows the organization to develop a wider range of internal representations of itself and its environment. This awareness allows agents of organizations to recognize and absorb complexity, thereby enabling responses that are appropriate and intelligent.

The importance of understanding and recognizing BI processes and mechanisms within enterprises and organizations cannot be overstated. They exist across a variety of organizations and enterprises (as the case study will prove) and are prevalent (IBM, 2011), as they support and complement information collection, integration, dissemination and decision-making, whether they are recognized under the umbrella term 'Business Intelligence' or not. However these processes and mechanisms are implemented, they remain an important aspect for consideration in any large organizations. They enable the 'catching on', 'making sense' of things, or 'figuring out' what to do next.

2.1.2 Business Intelligence and Competitive Intelligence

Business intelligence is a term that has been popularised by Gartner Inc. for a quarter of a century. It is used to refer to processes and systems used by organizations to systematically analyse and manage their internal activities, performance, capabilities and strategic goals in *competitive* environments (Bucher, Gericke & Sigg, 2009). BI has been described as a collection of technologies (Negash, 2004) that extend the capability, functionality and architecture of information systems (Russell, Haddad, Bruni & Granger, 2010) to provide knowledge workers with the ability to enhance and optimize their decision making capabilities (Chaudhuri, Dayal & Narasayy, 2011). It is regarded as a practice (Dayal, Castellanos, Simitsis & Wilkinson, 2009) or a methodology (Russell et al, 2010) that is implemented with the intention of reducing uncertainty and providing a platform for *awareness* and knowledge discovery. Mikroyannidis and Theodoulidis (2010) consider BI as a collection of techniques and tools that provide knowledge-support where it is required. All these competing perspectives from academia widen the precise meaning of business intelligence. Popovič, Turk & Jaklič (2010) point out that BI is often confused with software or technology components, while others, Arnott and Pervan (2005) in Popovič et al. (2010), regard this moving definition as an indication of vendors continually trying to reinvent their product ranges. This suggests that BI is not an academic term nor is it capable of being reduced down to a core set of functional capabilities. Koronis and Yeoh (2010) point out that a limited set of authoritive criteria for management reference exists for BI because the market is mainly driven by the IT industry and vendors. In spite of these misgivings, BI processes and practices have an important role to play in organizations: they must exist to provide knowledge workers with relevant and insightful information to enhance confidence in decision making.

When implemented correctly BI processes are able to deliver insights into organizational strengths and weaknesses, demonstrate opportunities and threats, and provide forecasts into future events (Watson & Wixom, 2007). The latest SIM survey of Kappelman et al. (2013) and international survey of Luftman et al. (2013) still place BI and analytics as the number one technology and application concern in industry. According to key decision makers BI was by far their single largest investment and area of concern. In a similar study in 2011 (in which more than 3000 CIOs participated), BI had become a massively important investment and was seen as a means to increase competitiveness over the next 3 to 5 years (IBM, 2011). While the nature of these BI implementations was not clearly documented, research in South Africa has indicated that many BI implementations tend to focus on retrospective reporting with little focus on prediction (Hart, 2009).

The introduction of Business Intelligence systems (BIS) is often coupled with a focus on improving information processes (Popovič et al, 2010). These include information quality improvement goals (e.g. improved self-service access to data), integration with various other data providers or sources, and interactive access to data. Alongside interpretation and analysis of data, these are considered first steps towards BIS investment justification.

Michalewicz, Schmidt, Michalewicz and Chiriac (2006) propose that the goal of BI is to collect, digest and present knowledge. This is achieved by (i) accessing data from multiple sources, (ii) transforming data into information and knowledge, (iii) combining and presenting this output in a graphical interface, and (iv) ensuring access to these interfaces and outputs. Watson and Wixom (2007) simplify this concept by describing it as a process made up of only two fundamental activities, namely data input and data output. The process of inputting data is more commonly referred to as data warehousing in which different data sources are integrated into a single coherent data warehouse. Traditionally data warehouse teams will combine information from various data sources and then enrich this data by transforming it. This data transformation process can include actions such as time-period aggregation, subject-orientation or score calculations and are commonly referred to as Extract, Transform, and Load or ETL (March & Hevner, 2007). The data input process is regarded as the most challenging aspect of BI and is estimated to require up to 80% of all time and effort of the entire BI cycle. Data sourcing and integration processes often require significant and complicated effort. This may be one of the significant factors affecting BI success (i.e. data quality, data ownership and legacy-system issues typically arise).

Chaudhuri et al. (2011) present a more technical architecture for BI across five focus areas. These are (i) data sources, (ii) data movement and streaming engines, (iii) data warehouse servers, (iv) middle-tier servers, (v) and front-end applications (see Figure 1).

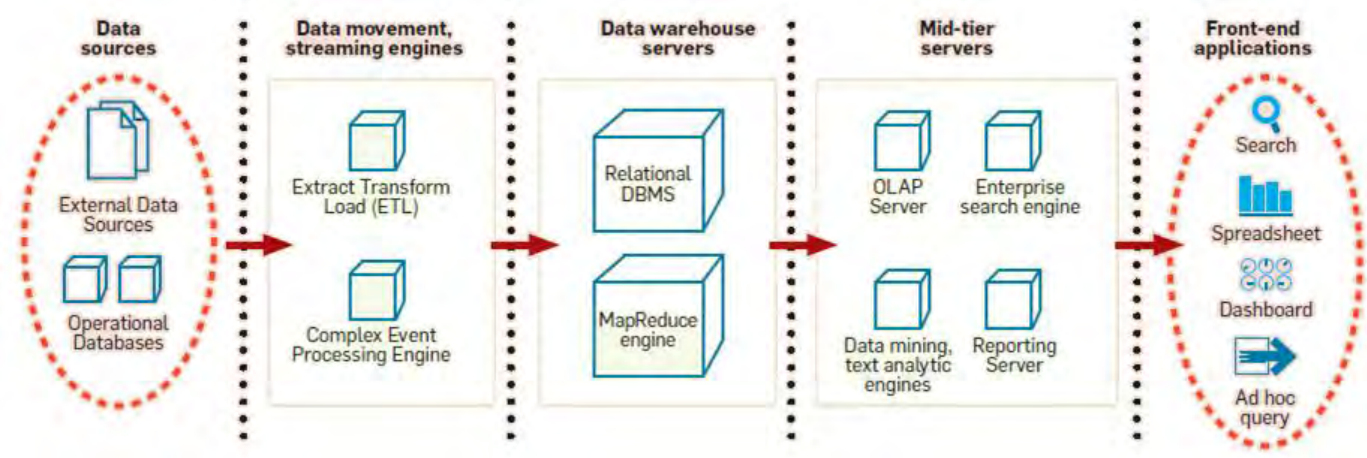


Figure 1. Five focus areas of a BI architecture (Chaudhuri, Dayal & Narasayy, 2011).

Zeng, Li and Duan (2012) describe a BI architecture that relies heavily on data mining as a core competency. Results are obtained through complex data mining processes and algorithms. Zeng et al. (2012) put forward the concept of the "analysis process of business intelligence", an iterative process used to solve business problems (see Figure 2). This process relies heavily on the application and use of data mining for extracting patterns and testing hypotheses and requires specialist skills (i.e. a data mining expert with knowledge of different algorithms to solve problems using appropriate techniques).

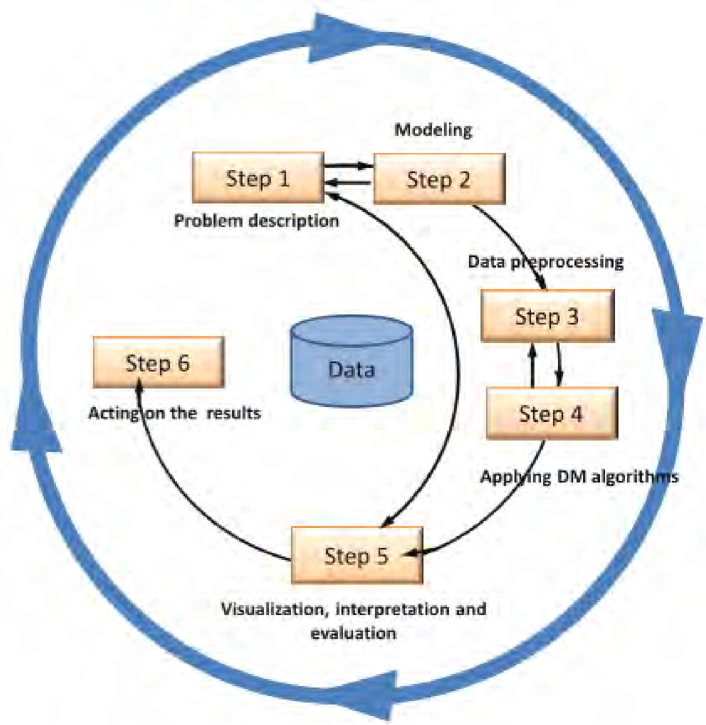


Figure 2 . Analysis process in applying business intelligence (Zeng, Li & Duan, 2012)

Traditional (business-driven) BI initiatives are focused on facilitating organisational decision-making by addressing the needs of business users (BU). This is largely realized by the corporate BI team who develop standard reports that contain monitored key performance indicators (KPI) in order to answer predetermined business questions (Yu, Lapouchnian & Deng, 2013). Pre-built reports tend to be used by managers to support daily operations or for planning purposes and can be generated by BUs. However, more and more often, BUs require additional ad-hoc reports from their BI teams in order to provide the right information at the right time. Because of the perceived value of analytics coupled with the explosive growth of data across enterprises, BI teams often struggle to deliver these ad-hoc reports in a timely fashion (see Figure 3).

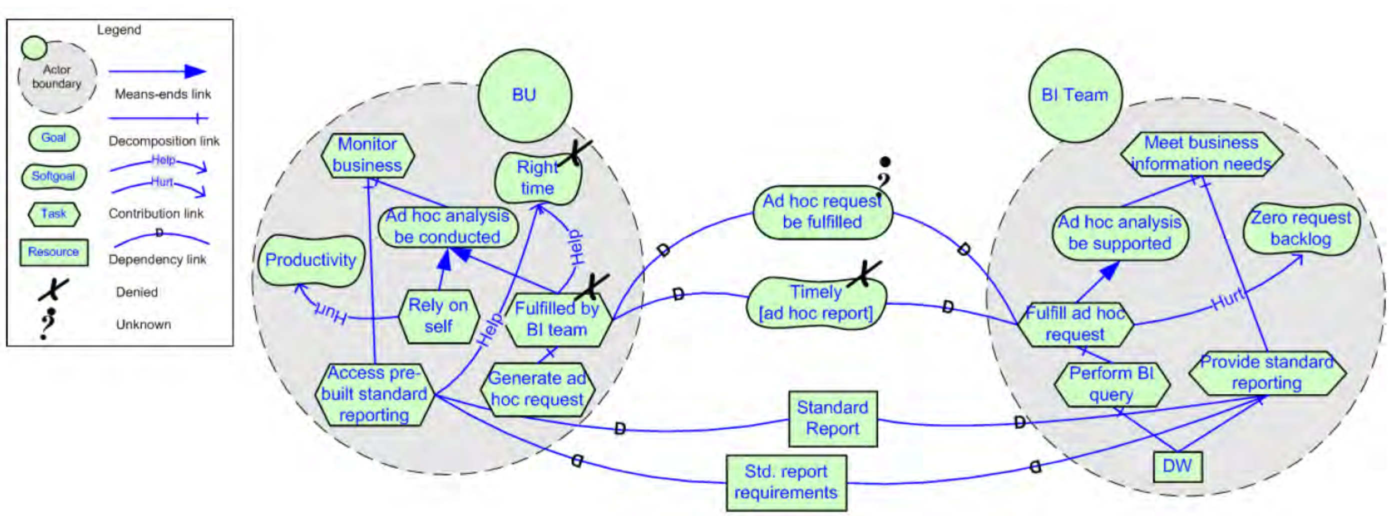


Figure 3. A social model for traditional BI with interaction taking place between Business Unit and the BI Team (Yu, Lapouchnian & Deng, 2013)

Corporate "traditional BI" processes are changing to cater for these new needs. Yu et al. (2013) describe the move to self-service business intelligence (SSBI) done in some organizations to address the reporting-delay created by overburdened BI teams (see Figure 4). This move gives business users freedom and shifts responsibility as they are expected to rely on themselves thereby removing dependency on others. Two different types of users have been identified who typically perform these self-service types of queries or analyses:

- *casual users* (which include executives, managers or front-line workers) are domain experts that utilize information to do their jobs; they need just-in-time analytics to make decisions within short timeframes.
- *power users* (which include data scientists, statisticians and business analysts) are typically hired to analyse information; they explore the potential value of data through various creative and iterative analyses; they have strong analytical proficiency with reasonable domain knowledge.

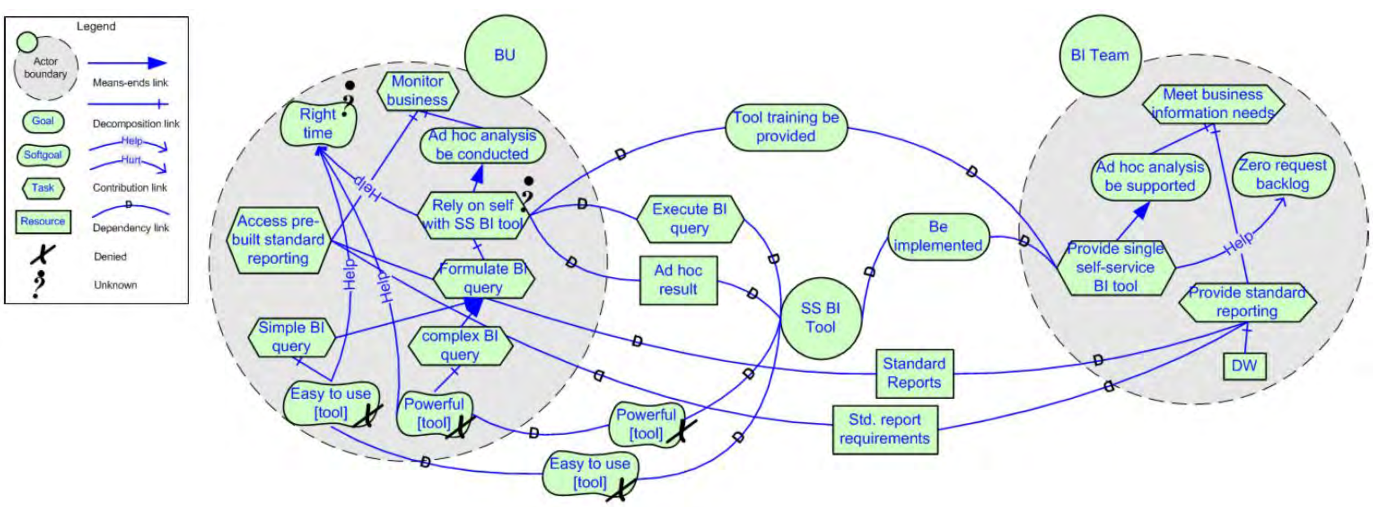


Figure 4. A social model for the move towards self-service BI (Yu, Lapouchnian & Deng, 2013)

Olszak and Ziemba (2007) state the purpose of BI systems is to address decision making concerns at tactical and operational levels. Tactical decision making would be to address the realisation of strategic objectives by optimising future actions (e.g. aspects of company performance at organisational, financial or technological levels). Operational decision making would be to address departmental related concerns around ongoing operations (e.g. up to date financials, coordination and cooperation with suppliers and customers, etc). These BI operations would support data analyses to address questions across different aspects of organisational performance. Examples include (Olszak & Ziemba, 2007):

a) *Financial analyses including*: costs and revenues; calculation and comparative analyses of income statements; analyses of balance-sheet and profitability; analyses of financial markets and sophisticated controlling;
b) *Marketing analyses including*: analyses of sales receipts; sales profitability and sales margins; meeting of sales targets; time of orders; actions of competitors; stock exchange quotations;
c) *Customer analyses including*: time of maintaining contact with customers; customer profitability; modelling customer behaviours and reactions; customer satisfaction;
d) *Production management analyses including*: Identifying production bottlenecks and delayed orders; Information on or understanding of production dynamics; comparative production reports per department or plant;
e) *Logistic analyses including*: utilizing efficient Information to identify partners of supply chain;
f) *Wage analyses including*: reports/information on employment types; payroll surcharges; personal contribution reports; analyses of average wages;
g) *Personal data analyses including*: employee turnover; employment types;

Competitive Intelligence (CI) is often confused with Business Intelligence as both describe information design and management processes with information products (providing insight). While BI processes are primarily focused on the organization and its internal activities and goals, CI processes are more focused on understanding and interpreting the external environment, i.e. competitive forces (Bose, 2008). CI is an important part of any organization's strategic planning and management process as it seeks to provide actionable intelligence and in doing so - a competitive edge to the organization (Kahaner, 1998 in Bose, 2008). This is achieved by sourcing data and information from outside of the organization to enable the anticipation of future events by predicting or forecasting changes in the macro-environment. CI practices are strongly focused on competitive forces or organizations that compete for the same resources but, similar to BI, CI has a strong emphasis on information gathering, integration, analysis and planning based on information obtained through a standardized process. The Society for Competitive Intelligence Professionals (SCIP) propose the CI process as a continuous cycle (see Figure 5) in which raw data is acquired, transmitted, evaluated, analysed and made available. This process consists of 5 phases (see Figure 5):

1) Planning and Direction: *define the information requirements of the organization* (what, why, when). This requires collaboration with decision makers to translate their needs into specific intelligence requirements. The outputs of this phase should provide purpose and direction for CI operations.
2) Collection: this phase includes the research and identification of all potentially relevant information sources in an ordered form.
3) Analysis: this phase calls for activities related to analysis of data such as the identification of patterns, relationships, or anomalies. This also requires a systematic examination of all data, information and knowledge for relevancy or significance to support decision making or the development of strategies. The output of this phase should be the recommendation of a specific action.
4) Dissemination: *report and inform*. This phase produces the finished product of previous phases and is communicated to decision makers in an easily understood format. Communication of these findings can be in the form of reports, dashboards or even meetings. Findings or recommendations are often used as inputs into further analysis (e.g. profiling, scenario planning, and scenario analysis).
5) Feedback: *evaluate*. This phase involves activities related to measuring the impact of the intelligence provided (what, why, how). This phase is important in providing analysts with areas for continuous improvement or further investigation.

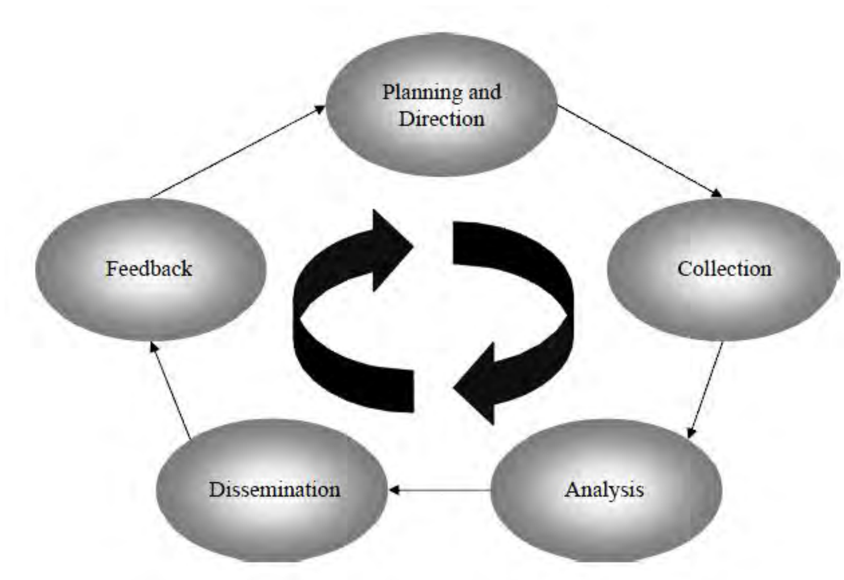


Figure 5. The five phase cycle of the CI process (Bose, 2008).

While CI has a different focus to BI (i.e. the context or macro-environment), the researcher felt its inclusion was necessary as the 5 phase lifecycle provides clear guidance on the use of information for action. While BI's focus appears to be more localised with emphasis on the management and prediction of changes and conditions internal to the organization, the CI perspective also takes into account the external environment with a focus on understanding and predicting its change. Together these two disciplines may support one another in the storage and use of information for better planning, action and outcome.

2.1.3 BI Focus Areas and Maturity

Russell et al. (2010) describe the evolutionary 'steps' that organizations may experience during their realization or investment in business intelligence processes or mechanisms. These 'steps' or 'chasms' may be experienced as a result of internal or external influences. While BI implementations are ongoing processes extending well beyond their initial implementation it is suggested that this evolution is triggered by the presence of social, environmental and technological shifts. Gartner's Business Intelligence and Performance Management Maturity model (BIPMM), see Figure 6, focuses on organizational characteristics and change, suggesting that BI and performance management will change over time.

Figure 6. Gartner's Business Intelligence and Performance Management Maturity model (Russell, Haddad, Bruni & Granger, 2010).

Tan, Sim and Yeoh (2011) draw out four major areas of focus for a business intelligence programme (summarised in Table 1) and synthesise a maturity path for enterprises across four focus areas. Information Quality (IQ), Master Data Management (MDM), Warehousing Architecture and Analytics form the basis of this maturity model across five levels of maturity which can be used to benchmark an enterprise's BI initiatives. This model also assists the organization or enterprise to better plan, assess and manage their activities towards the improvement of BI processes and mechanisms. In doing so this model provides a roadmap for enterprises and organizations. The authors also go into more detail describing the four focus areas of a BI programme at each of the five levels of maturity.

Level	INFORMATION QUALITY	MASTER DATA MANAGEMENT	WAREHOUSING ARCHITECTURE	ANALYTICS
5	**SINGLE VIEW OF TRUTH** Source of information quality problems have been recognised. There are continuous initiatives to improve processing of information quality problems. Besides, impact of poor information quality has been calculated.	**ENTERPRISE DATA CONVERGENCE** In this stage, the hub is fully integrated into the application system environment. The hub will propagate data changes to all the application systems that need the master data. Application processing occur without depending on physical system location and data navigation.	**ANALYTICAL SERVICES** Gradually, the enterprise data warehouse value increases as its visibility declines. Enterprise data warehouse fades into the background as a business intelligence service. Examples of analytical services are interactive extranets, web Services, decision engines and so forth.	**ANALYTICAL COMPETITOR** The enterprise-wide analytics capability promises the company regular benefits. The company focuses on continuous analytics review and enhancement.
4	**IQ ASSESSMENT** Information quality metrics have been developed and information quality is being evaluated.	**BUSINESS RULE & POLICY SUPPORT** A process-driven data governance framework exists to maintain centralized business rules management and distributed rules processing. Organisation has a mature change management process. SOA is applied to integrate common business methods and data across applications. There is an automated way to both enforce and undo changes to master reference data.	**ENTERPRISE DATA WAREHOUSE** Enterprise data warehouse acts as an integration machine that continuously merges all other analytic structures into itself. The enterprise data warehouse helps organisation to achieve a single version of the truth.	**ANALYTICAL COMPANY** Analytic capability draws most attention from company top executives, thus enterprise-wide analytics capability is being developed.
3	**IQM INITIATIVE** In this stage, information quality management is treated as a core business activity and widely implemented across organisation.	**CENTRALIZED HUB PROCESSING** In brief, everything is centralized during this stage. Master reference data, business- oriented data rules, and connected processing are centrally handled. Cross-functional or cross-organisation conflict can be resolved by a data governance process. Thus, data accuracy and consistency is guaranteed.	**DATA WAREHOUSE** A data warehouse provides interactive reporting and deeper analysis. New insights are promised due to the capability of cross-functional boundaries query.	**ANALYTICAL ASPIRATIONS** Executives commit to analytics by aligning resources and setting a timetable to build a broad analytical capability.
2	**DEFINE IP AND IQ** All Information Product (IP) and Information Quality (IQ) requirements have been identified and documented. Accordingly, related information quality dimensions and requirements have been classified.	**PEER-BASED ACCESS** There is hardcoded logic for applications to interact with the list of master data. A data model is created to identify each master record distinctively. Individual applications take responsibility to maintain the master list. All data and integrity rules are copied to new integrated application systems.	**DATA MARTS** A data mart is an analytical data store that generally focuses on specific business function within an organisation, e.g. department. Data marts are tailored to meet the needs of data users. Usually interactive reporting tools such OLAP and ad hoc query tool are used to access the data marts to gain deeper insight.	**LOCALIZED ANALYTICS** Functional management builds analytics momentum and executives' interest through applications of basic analytics.
1	**AD-HOC** Information Management (IM)/Information Quality Management (IQM) processes are not standardized or documented during this stage. There is no awareness of any information quality (IQ) issues, therefore no attempts are made to assess or improve information quality. Organisation acts in response only when information quality problems occur.	**LIST PROVISIONING** There is no systematic and thorough way of ensuring changes to the master list. Defining and maintaining master lists involve significant meetings and human interaction. Data conflicts, deletions, changes, explaining data file formats, and content details are handled manually. Individual applications must understand how to navigate to the master list.	**SPREAD MARTS & MGMT REPORTING** Management reports are static reports which are printed and disseminated to employees on weekly, monthly, or quarterly. Spread marts are spread sheets or desktop databases that function as surrogate data marts.	**ANALYTICALLY IMPAIRED** The company has some data and management interest in analytics.

Table 1. Five levels of maturity across four focus areas for Enterprise Business Intelligence (Tan, Sim & Yeoh, 2011)

2.1.3.1 Data/Information Quality

Data and information are often used interchangeably, as are the terms information quality and data quality (Baškarada & Koronios, 2013). For the purpose of this research they shall be referred to as the latter. Data quality (DQ) has been an area of research since the 1980's and since the 1990's academic institutions have focused on management programmes and frameworks to address quality improvements in business. Data is considered of high quality when it is fit for intended use in operations, decision making and planning (Lucas, 2011). In the corporate domain challenges around DQ are heavily influenced by inter-data models or data architectures which affect data integration and consistency.

Lucas (2011) describes the measurement of DQ across two perspectives, namely "depth" and "width" (see Figure 7). As a multidimensional concept DQ "depth" is operationalized through dimensions which are the data characteristics valued by its consumers - namely *accuracy*, *timeliness*, *interpretability*, *completeness* and *relevancy*. DQ "depth" therefore refers to the DQ perspective valued by data consumers within a single sub-system as measured by data-characteristics but when data integration occurs across sub-systems to create holistic views of business the "width" perspective becomes the narrative. This "width" can be thought of as a characteristic of data consistency across sub-systems and is commonly referred to as "master data management" (addressed under MDM section 2.1.3.2).

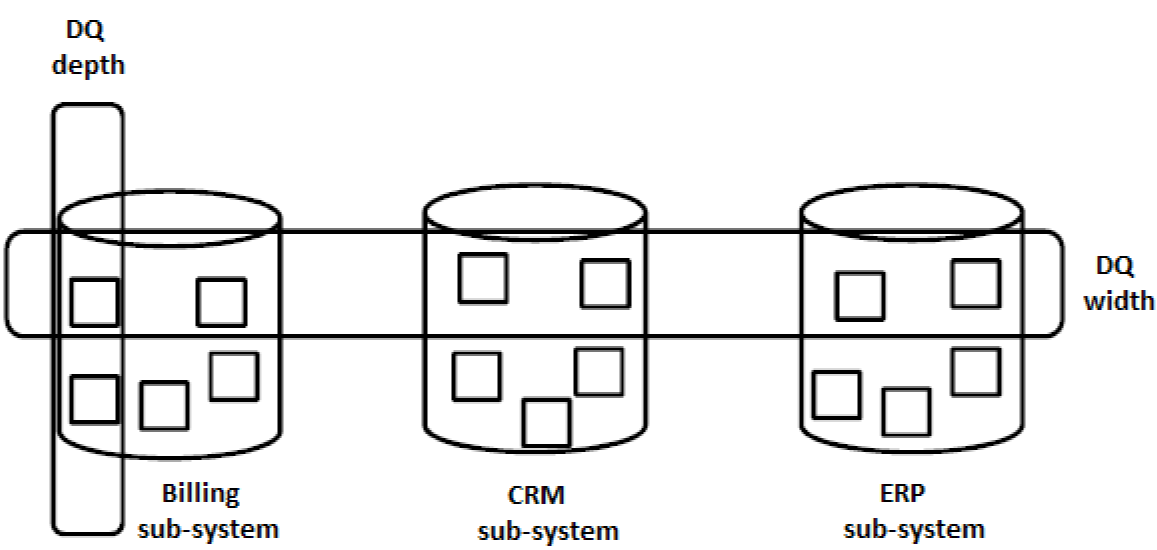


Figure 7. Corporate Data Quality "width" and "depth" perspectives (Lucas, 2011)

Weiskopf and Weng (2013) conduct a review of data quality in electronic health record (EHR) systems and provide a different set of dimensions and methods.

- *Completeness*: Is a truth about a (patient) record present in the system?
- *Correctness*: Is an element that is present in the system true?
- *Concordance*: Is there agreement between elements in the system, or between the system and another data source?
- *Plausibility*: Does an element in the system make sense in light of other knowledge about what that element is measuring?
- *Currency*: Is an element in the system a relevant representation of the (patient) record state at a given point in time?

A deeper view into the terms used to denote each of these dimension is as follows (Table 2):

Completeness	Correctness	Concordance	Plausibility	Currency
Accessibility	Accuracy	Agreement	Accuracy	Recency
Accuracy	Corrections made	Consistency	Believability	Timeliness
Availability	Errors	Reliability	Trust worthiness	
Missingness	Misleading	Variation	Validity	
Omission	Positive predictive value			
Presence	Quality			
Quality	Validity			
Rate of recording				
Sensitivity				
Validity				

Table 2. Terms used to describe five common dimensions of data quality (Weiskopf & Weng, 2013)

Weiskopf and Weng (2013) identify seven core categories of DQ methods to assess the different DQ dimensions. These general methods are listed in descending order of relevance (see Figure 8) and are based on the findings from their study :

- *Gold standard*: requires data sets drawn from multiple sources to verify and test for completeness and correctness.
- *Data element agreement*: two or more elements within a system are compared to test that they report the same or compatible information.
- *Element presence*: used to determine if desired or expected data elements are present.
- *Data source agreement:* data is compared with other sources to determine if they are in agreement.
- *Distribution comparison*: summary statistics or distributions of aggregated data are compared with expected distributions for concepts of interest.
- *Validity check*: data is assessed using a variety of techniques to determine if values "make sense".
- *Log review*: information on data entry practices are examined (e.g. dates, times, edits, etc)

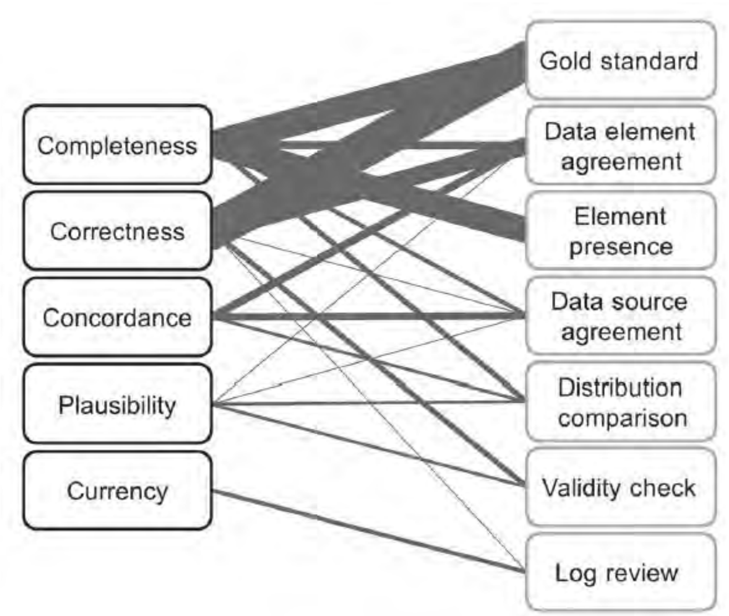


Figure 8 Mappings between DQ dimensions and assessment methods with pairings weighted according to (literature) relevance (Weiskopf & Weng, 2013)

2.1.3.2 Master Data Management (MDM)

Master data falls under a special category of enterprise level data that refers to the core 'nouns' within a specific business context. The processes and systems designed to maintain this data throughout its life-cycle are regarded as master data management or MDM (Bai, Li, Li & Song, 2010). Some maintain that MDM is the realisation of the need to enhance organization wide data quality ascribing it to the "width" perspective of data quality (Lucas, 2011). Nonetheless it offers opportunities for technology and business process improvement.

In order to understand MDM better a data taxonomy can be used to distinguish the different 'types' of data found in an enterprise. Tozer (1999) in Cleven and Wortmann (2010) divides enterprise-level data into two major categories, namely Domain data and Metadata (p. 1). These categories are further subdivided into Master data, Transactional data, Informational metadata and Operational metadata (see Figure 9).

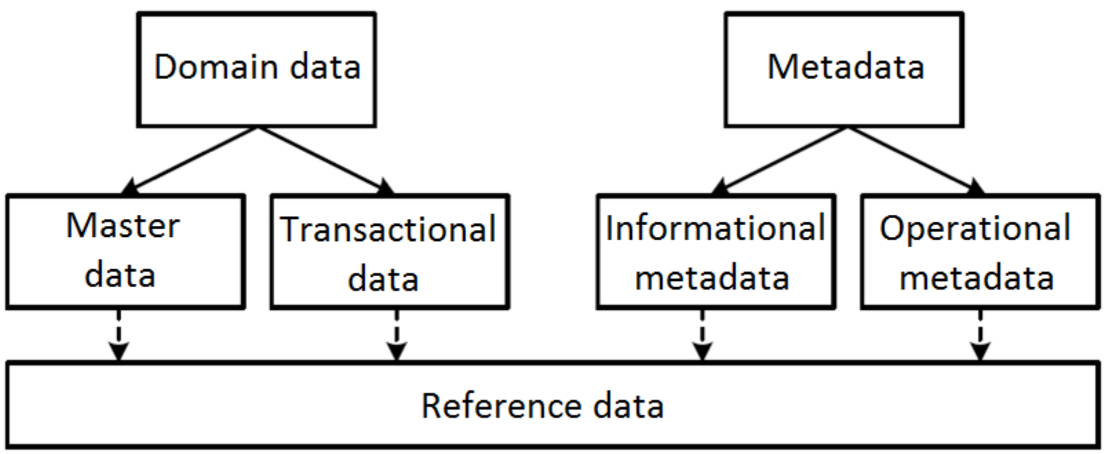


Figure 9. Data Taxonomy (Cleven & Wortmann, 2010)

Domain data refers to subject-area data (e.g. a live information system). The *master data* class refers to the common entities upon which an organisation's processes and systems make repeated references. This class of data is sometimes referred to as *persistent data* (Cosma, Văleanu, Cosma, Vasilescu &

Moldovan, 2013). Classic examples include service-points, products and customers. Master data objects have particular characteristics (Cleven & Wortmann, 2010):

i) *Existential independence:* master data objects can exist on their own (e.g. a health facility, doctor, nurse or patient can exist independently of one another while a script or drug invoice cannot);

ii) *Low change frequency:* these objects are stable throughout their life-cycle with changes to their attributes being uncommon;

iii) *Volume stability:* these objects rarely increase in number, unlike sales orders, purchases or transactions;

Transactional data represents the data generated by common business processes and it may change many times during its lifecycle because of its intended design and nature. This class of data is sometimes referred to as *operational data* (Cosma et al, 2013). Transactional data makes repeated reference to master data within a domain or system. Transactional data objects tend to have the following characteristics (Otto & Reichert, 2010):

i) Usually *Time referenced* by a time dimension (e.g. items may be in stock at a particular time or during a period of time);

ii) *Modification frequency* is high (e.g. sales-order changes, transaction modifications, business-process status updates, etc)

iii) *Volume instability* as transactions will increase or decrease following business activity;

iv) *Existential dependence* because transactional data cannot exist independent of master data;

Operational metadata is technical data describing the design and operation of information systems while *Informational metadata* is used to support understanding and access to domain data for end-users.

Cleven and Wortmann (2010) group all four of these data categories into a set of *reference data* or an agreed-upon range of values. Cleven and Wortmann (2010) go further in identifying three core master data domains (see Figure 10). *Party* refers to business-relationship entities and commonly includes people or organizations (e.g. supplier, distributor, employee or customer). *Thing* can refer to a product, service or asset that is offered or owned by organization and is usually driven or distinguished by industries and their product characteristics. *Location* refers to places, regions or sites and is often used together with *party* or *thing* to determine where products are sold or produced.

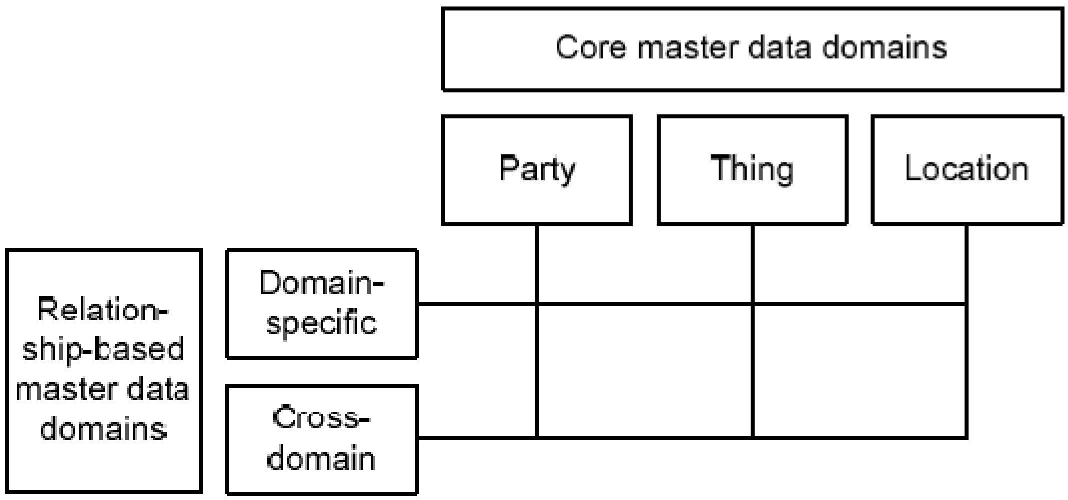


Figure 10. Different master data domains (Cleven & Wortmann, 2010)

Dreibelbis, Hechler, Milman, Oberhofer, van Run and Wolfson (2008) in Cleven and Wortmann (2010) propose two additional relationship-based master data domains, namely *domain-specific groupings* (e.g. categories of services, hierarchies, etc) and *cross-domain relationships* (e.g. product 'x' is produced by supplier 'z') which to different things, locations and parties to one another (p. 2). These relationship-based master data domains appear more suitable for use in analytics and reporting.

MDM is regarded as an application-independent process that provides guidelines for the management of this data which may or may not be hosted in its own data source (Otto & Reichert, 2010). The goal of MDM is to "provide organizations with the ability to integrate, analyse and exploit the value of their data assets, regardless of where that information was collected" (Cleven & Wortmann, 2010, p. 1). Otto and Reichert (2010) make reference to different MDM focus areas that stem from the practitioners' domain. These include "*understanding master data integration needs*", "*defining and maintaining the data integration architecture*" and "*managing changes to master data*".

Cleven and Wortmann (2010) present a MDM implementation model comprised of elements that organizations can use to avoid common problems such as operational malfunctions, ineffective decision-making, time wastage and unnecessary (human) resource expenditure. Their model requires an approach that addresses both technical and organizational aspects. Backed up by Smith and McKeen (2008) and Berson and Dubov (2007), Cleven and Wortmann (2010) theorize an approach that requires the establishment of a supportive organization with adequate processes, emphasising "organizational preparedness" as a key issue. This "preparedness" requires the configuration of five fundamental components: master data structure, master data systems architecture, master data governance, master data processes and master data quality (see Figure 11).

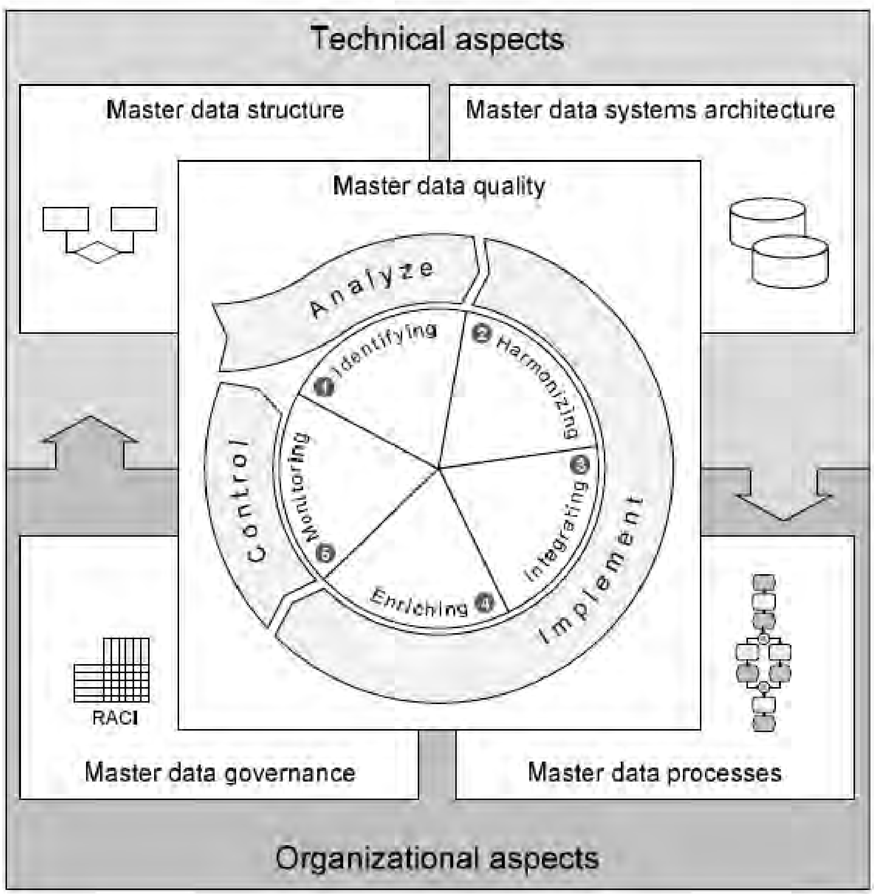


Figure 11. Core elements of Master Data Management (Cleven & Wortmann, 2010)

Configuring a ***master data structure*** requires an agreed-upon understanding of each data-domain's object definition as well as an understanding and model of object relationships. The ***master data***

systems architecture deals with the design of systems to support the different steps within each objects master data life cycle. This generally includes creation, storage, access, archiving and retirement processes. This component can be addressed through different approaches, one of which includes the development of definitions for *master data standards* for each system to comply with.

Establishing the ***master data governance*** component requires a clearly articulated mission statement and the formation of organizational support structures. These structures require clearly defined roles, activities and responsibilities. ***Master data processes*** are designed to prescribe organizational counterparts within the systems architecture with regards to activities and tasks such as creating, using, maintaining and archiving master data objects. Very importantly this component specifies how communications, support and training is to be conducted.

The process of data quality management (DQM) is based on total quality management (TQM) concepts and practices. It is accomplished by developing and implementing data quality policies and guidelines alongside active DQ assessment processes which includes activities and focus areas such as *data quality auditing* and *certification*, *data quality analysis*, *data cleansing*, *data correction*, *data quality process improvement*, and *data quality education* (Lucas, 2011). This approach considers the management of data as a “corporate asset” and is arranged into comprehensive set of concepts, roles and responsibilities (see Table 3). These concepts, roles and responsibilities extend across the “width” perspective of DQ in organizations and predominantly address the administrative application of DQ.

Concept/Role or Responsibility	Definition
Data Policy (**DP**)	A **DP** is a statement that delineates management responsibility for data and activities that touch and/or impact data and information. It can cover the following categories: – Data Quality in its broadest sense; – Data assets inventory; – Data sharing and availability; – Data architecture; – Data security, privacy and appropriate use; – Data planning.
Data Governance (**DG**)	**DG** refers to the framework for decision rights and accountabilities to encourage desirable behaviour in the use of data. To promote desirable behaviour, data governance develops and implements corporate-wide data policies, guidelines, and standards that are consistent with the organization's mission, strategy, values, norms, and culture. – **DG** is also the exercise of authority, control and shared decision-making (planning, monitoring and enforcement) over the management of data assets.
Data Owner (**DO**)	**DO** is the entity (usually a business unit) having responsibility and authority for a specific dataset. Although this definition is not consensual, it is the one we found in the case study environments.
Data Steward (**DS**)	**DS** is a business leader and/or subject matter expert designated as accountable for: – The identification of operational and business intelligence data requirements within an assigned subject area; – The quality of data names, business definitions and domain values within an assigned subject area; – Compliance with regulatory requirements and conformance to internal data policies and data standards; – Application of appropriate security controls; – Analysis and improving of data quality; – Identification and solution of data related issues. They should also be considered data subject-matter experts for their respective business functions and processes

Data Quality Champion (**DQC**)	**DQC** is a manager who actively and vigorously promotes their personal vision for using data quality related technology innovations. They push projects over approval, provide political support, keep participants informed, and allocate resources to data quality projects.
Data Quality Assurance (**DQA**)	**DQA** is the part of data quality management focused on providing confidence that quality requirements will be fulfilled
Data Quality Control (**DQC**)	**DQC** is the part of data quality management focused on fulfilling quality requirements
Data Quality Methodology (**DQm**)	– A **DQm** is “a set of guidelines and techniques that, starting from input information describing a given application context, defines a rational process to assess and improve the quality of data”. – A **DQm** is made of phases and activities and uses techniques (**DQT**) and tools (**DQt**) to accomplish its work.
Data Quality Technicians (**DQT**)	**DQTs** can be data and process driven: – Data driven **DQTs** correspond to algorithms, heuristics, knowledge-based and learning processes that provide a solution for specific DQ problems, like record linkage (eg finding and merging duplicates, i.e. different records that represent the same real world entity), standardization techniques (comparing data with lookup tables, and updating it accordingly) or data and schema integration; – Process driven **DQTs** are used to describe, analyse and reengineer the information production processes.
Data Quality Tools (**DQt**)	**DQts** are software products that implement specific **DQTs** to address the core functional requirements of the data quality discipline, in particular profiling, parsing and standardization, generalized "cleansing”, matching, monitoring and enrichment.

Table 3. Main concepts, roles and responsibilities for the data management approach (Lucas, 2011)

Master data quality improvement is a continuous and ongoing endeavour across organizations and is the fifth component of MDM according to Cleven and Wortmann (2010. It is based on strategic management and is a three phase process: *analyse*, *implement* and *control*. The *analysis* phase is used to identify and assess the current picture of organizational data objects. This is followed by the *implementation* phase which involves three different activities (*harmonizing, integrating* and *enriching)*. Semantic *harmonization* must be developed for each of the core data entities, followed by data *integration* which typically involves the matching, normalizing, cleaning and synchronizing of master data from different sources in the organization. Thereafter it is common for master data *enriching* to take place by adding additional metadata (organizational and technical) and sometimes even external data to enhance value. The third and final phase relates to (quality) *control*. Achieved improvements are protected by preventing the introduction of data errors. This is accomplished by using defined standards, policies, metrics and performance indicators (Cleven & Wortmann, 2010).

2.1.3.3 Data Warehousing (DW)

The data warehouse concept was introduced in 1992 by W.H. Inmon as a special database for managing large volumes of data to support decision making. Data from various sources (e.g. operational and external databases) are built up and combined into special data of a reasonable size in order to serve as support for decision makers. Cosma et al. (2013, p. 370) provide the following description:

"Data warehouses are non-volatile data collections, oriented to the subject, integrated, variable in time and supporting the managerial decision making process".

A simplified architecture for the use of data warehousing in organizations to support decision making is presented (see Figure 12) in which three zones or functional spaces are described. The *data source* zone comprises source data systems, most often provided by operational or transactional database systems, which utilizes the ECTL process (extraction, cleaning, transformation and loading) to integrate data into the *data warehouse* zone. This data warehouse zone typically makes use of a uniform database structure in which data cannot be modified. It is generally comprised of *(heavily or slightly) Aggregated data*, *Detailed data* and *Meta data*. The distribution of data from the warehouse into the third zone (*access instruments and OLAP use* area) is accomplished by views across different data marts sometimes split into subject areas. In this final zone, specialized data processing instruments and technologies should be present.

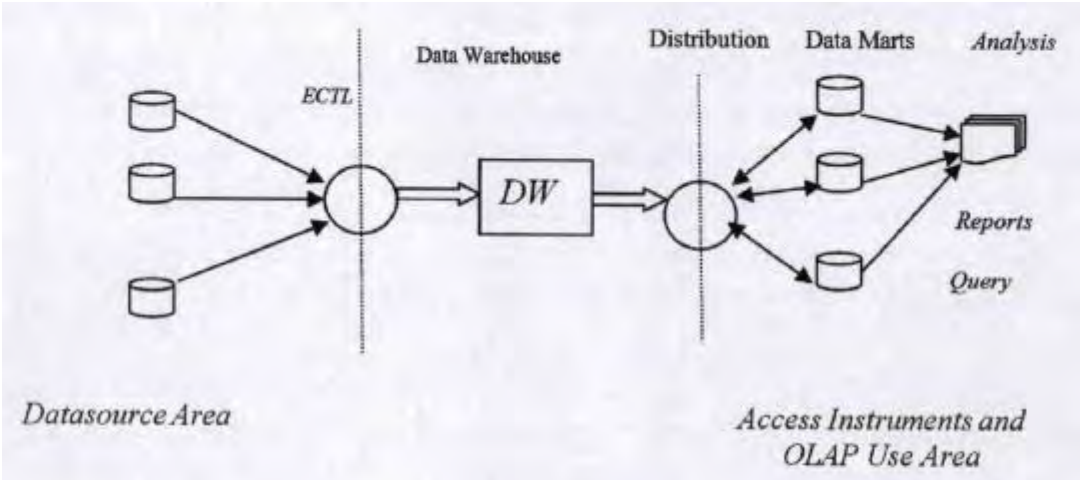


Figure 12. Three functional areas for data warehousing in organizations (Cosma, Văleanu, Cosma, Vasilescu & Moldovan, 2013)

Data warehouse schemes are the sum of their data marts, data sources and the profiles required by their users. They make use of multidimensional data models to store (meta-data) attribute trees, specific (data) dimensions, (data) measurements and hierarchies. The design of a DW requires an understanding of source database (system) attributes (e.g. dimensions, measurements and hierarchies) as they are defined together with users (Cosma et al, 2013).

Multidimensional concept models are considered to be the foundation for data warehouse designs (Gosain, Nagpal & Sabharwal, 2011) with multidimensional data allowing us to harness the power and capabilities of analytics (Cuzzocrea, Davis & Song, 2011). Examples of multidimensional data models include *multidimensional abstractions*, *hierarchy-based dimensional tables*, *multi-resolution fact tables* and *multi-way aggregations* which help us achieve more powerful analytics by enhancing and adapting existing models.

Inmon, Strauss and Neushloss (2010) describe the some of the challenges of building a data warehouse. These include data integration, data volumes and development approaches. Data integration processes involving legacy systems are tricky. Legacy systems tend to be intractable, difficult to change and unaccommodating which works against the natural process of data integration. Legacy data tends to be application-oriented and the conversion to corporate data usually requires changes to source systems which are never easy and almost impossible within legacy environments. Data growth within warehouses is a known issue for most IT professionals. It is commonly accepted and good practice to discard old data within application systems as this data is often undesirable and may even slow down operational environments. However not all historic data is undesirable. For some analyses historic data is indispensable and having a convenient place to store this data is recommended. The DW development approach does not follow the traditional software development process of requirements gathering and then building. A DW is built up incrementally over a series of iterations and the process should cater for discovery and learning. It is impossible to know ahead of time what end users will require from their data.

Inmon et al. (2010) present the data warehousing 2.0 environment hinting at the need for information restructuring across the enterprise so as to accommodate corporate data integration (pp. 10). This approach requires a change from application-oriented design towards a corporate landscape structure (which seems to support the need for data quality "width" or master data management).

Kimball and Ross (2011) describe the components of the data warehouse according to categories of data. Anything that does not qualify as actual data is considered metadata. These include administration and user groups, schemas, data that guides transformation processes, cleansing rules, conformed dimension and fact definitions, schedules, log results, etc, which forms part of the data warehouse metadata framework. Their position is that an over concern regarding the categorization and management of this metadata distracts efforts aimed at building dimensional models.

2.1.3.4 Analytics

Data analytics has a long history in the area of database management and has always been strongly reliant on data collection, extraction and analysis processes. It is referred to as the techniques, technologies, systems, practices, methodologies and applications used to (critically) analyse business data in order to help enterprises understand their business and market to support effective decision making (Chen, Chiang & Storey, 2012). Analytics can also be defined as the complex procedures running against large data repositories with the goal to extract useful knowledge (Cuzzocrea et al, 2011).

Data management and data warehousing are considered the foundations of analytics in enterprises while data mart designs and tools for ETL are essential for data conversion and integration. Chen et al. (2012) refer to business intelligence and analytics (BI&A) as a unified term with various capabilities. Their capabilities are inclined to focus on structured data and have a provenance based predominantly on statistical methods developed in the 1970's, data mining techniques developed in the 1980's and analytical techniques popularized in the 1990's.

The area of analytics that is heavily focused on database management solution (DBMS) and structured content falls under the space known as BI&A 1.0 (summarised in Table 4) and is usually coupled with a variety of capabilities but Gartner (Sallam, Richardson, Hagerty & Hostmann, 2011) consider eight as essential for analytics 1.0: *reporting*, *dashboards*, *ad hoc querying*, *search-based BI*, *OLAP*, *interactive visualization*, *scorecards*, *predictive modelling* and *data mining*. Business performance management (BPM) seems to have popularized the use of *score cards* which typically accompany analytics 1.0 (Chen et al, 2012) while *dashboards* are more frequently used to analyse and demonstrate performance metrics. Examples of statistical analysis and data mining techniques that are frequently used include *association analysis*, *data segmentation*, *clustering*, *classification*, *regression analysis*, *anomaly detection* and *predictive modelling*.

With the rise of the internet web-based offerings created a unique opportunity for analytics with a focus more on unstructured data. This area, known as BI&A 2.0, is dependent on gathering and analysing (web based) textual data using complex techniques (Chen et al, 2012) borrowed and adapted from BI&A 1.0. Examples include *web mining*, *social network analysis* and *text mining*.

In 2011, for the first time in history, mobile devices outnumbered laptops and PCs. This has created an ecosystem that generates data across a variety of mobile devices and includes (sensor-based) internet-enabled devices creating new opportunities for mobile analytics – which currently form the third (final) area of analytics known as BI&A 3.0 (Chen et al, 2012). These mobile platforms offer opportunities for analytics that are location-aware, person-centred and context-relevant but the underlying techniques for collecting, processing, analysing and visualizing all this data is still developing.

BI&A is highly data-driven and numerous opportunities exist to make use of abundantly available data. If applied correctly to a specific domain analytics could be used in many critical and high-impact application-areas. This would require an understanding of the various applications that collect data along with their data characteristics. In this way researchers and practitioners would be able to design appropriate analytic techniques that derive the intended impacts (Chen et al, 2012).

BI&A	KEY CHARACTERISTICS	GARTNER BI PLATFORMS CORE CAPABILITIES	GARTNER HYPE CYCLE
1.0	**DBMS-based, structured content** • RDBMS & data warehousing • ETL & OLAP • Dashboards & scorecards • Data mining & statistical analysis	• Ad hoc query & search-based BI • Reporting, dashboards & scorecards • OLAP • Interactive visualization • Predictive modelling & data mining	• Column-based DBMS • In-memory DBMS • Real-time decision • Data mining workbenches
2.0	**Web-based, unstructured content** • Information retrieval and extraction • Opinion mining • Question answering • Web analytics and web intelligence • Social media analytics • Social network analysis • Spatial-temporal analysis		• Information semantic services • Natural language question answering • Content & text analytics
3.0	**Mobile and sensor-based content** • Location-aware analysis • Person-centered analysis • Context-relevant analysis • Mobile visualization & HCI		• Mobile BI

Table 4. BI&A Evolution: Key Characteristics and Capabilities (Chen, Chiang & Storey, 2012)

Andrienko and Andrienko (2013) discuss visual analytics as a multiplier of analytic power. Research into this area focuses on finding effective methods that combine visual techniques with algorithms for data analysis. The goal is to design techniques where “visualization and computation interplay and complement each other” (Andrienko & Andrienko, 2013, p. 56). Much emphasis is placed on numeric time series (TS) data where values are attributed to different locations in space and time. TS analysis is well-established as an area within statistics and they (Andrienko and Andrienko) utilize clustering and interactive grouping to explore relationships between locations and objects.

2.2. Complex Adaptive Systems Theory

CAS theory has its origins in systems thinking (Hammer, Edwards & Tapinos, 2012) and is strongly influenced by complexity theory (Anderson, 1999). Merali, Papadopoulos and Nadkarni (2012) describe a complex system as being made up of many components that are embodied by resources, capacities to act and the potential to be connected in a variety of ways. Variations in agents or connections allow for diverse relationships to exist, resulting in diverse exchanges or feedback loops (Merali et al, 2012). As a result of this diversity and the natural ability for agents to self-organize and exchange resources, small inputs can have dramatic and unpredictable effects on system outputs (Mills, Rorty & Werhane, 2003; Nan, 2011).

2.2.1 CAS Characteristics and Adaptation

Complexity has long been considered a structural characteristic within organizations and their environments (Anderson, 1999) and the complexity in a CAS is derived from the partially connected nature of the system and its components. This makes behaviour within a CAS difficult to predict. Shaw (2009) describes a CAS as being comprised of its parts, the behaviour of those parts and the emergent behaviours of the system as a whole. A CAS is able to survive and adapt to external and internal environmental-state changes by relying on its underlying components and the relationships between these components within its environment (Meso & Jain, 2006). These characteristics allow the CAS to become self-organized and adapt to changing environmental conditions while retaining its integrity and identity (Merali et al, 2012).

While there may be many aspects to CAS theory, it has long been recognized by its three predominant components: *agents*, *interactions* and its *environment* (see Figure 13). In fact the most active elements of a CAS are considered to be those “adaptive agents” that interact in the environment. This is in reference to individuals, groups or coalitions of groups (Shaw, 2009). Those agents that change by adaptation do so by adapting their ‘rules’ as their experiences accumulate (Nan, 2011).

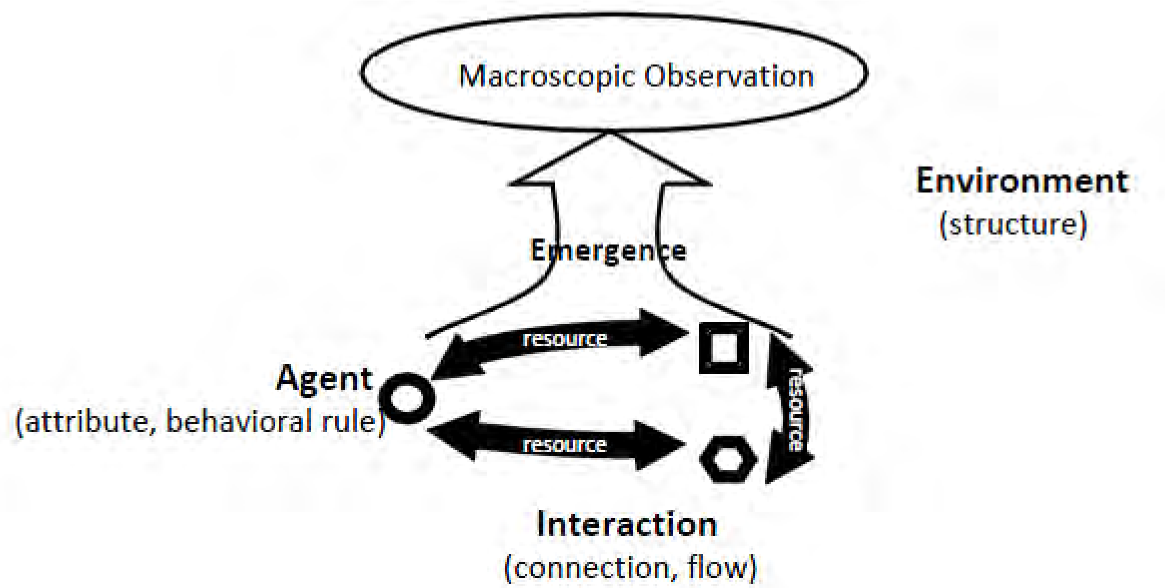


Figure 13. A CAS theory representation of its common themes (Nan, 2011).

One important feature of a CAS is its ability to organize itself into its own hierarchical arrangements as this allows the system and its related sub-systems to operate independently and interdependently of each another. This behaviour facilitates independence and self-adaptation that would otherwise be absent in the presence of a centralized controller. In place of a centralized controller Sturmberg, O'Halloran and Martin (2012) propose the existence of *shared values* or an *attractor* around which interactions and activities are focused. These *shared values* or *attractors* will be present throughout the system, existing across different levels of hierarchical arrangements supporting interdependency between sub-systems.

Meso and Jain (2006) state that CAS underpins the idea that there is a dynamic interplay between people, processes and products and that agile methods inform this interplay. They reinforce the idea that agility (in reference to adaptation) and complex adaptive systems coexist and support one another. They describe principles of a CAS as follows:

i) **Open Systems:** components of the system interact and exchange information or energy with the environment while operating in conditions far from equilibrium.

ii) **Interactions and Relationships***:* agents of the system interact dynamically to exchange information or resources with each other. The effects of these exchanges are propagated throughout the system even when exchanges are limited to only a few. The behaviour of the system is determined by the nature of interactions and not by the attributes of agents.

iii) **Transformative Feedback Loops**: interactions between agents result in direct and indirect feedback loops, i.e. changes to one part of the system are propagated via boundary maintaining rules to other parts of the system causing recipients-of-information to react or change in some way. These secondary reactions or changes result in feedback being transmitted back to the originator, causing a similar reaction. These propagations have the ability to support continuous improvement and create a sense of shared ownership.

iv) **Emergent Behavior**: the behavior of a CAS is unpredictable. Interactions between components may appear random but are often rich and dynamic in nature. As a result a CAS will demonstrate emergent and unanticipated behavior based on these interactions. Novel concepts or outputs may emerge from these interactions.

v) **Distributed Control**: a CAS cannot thrive in the presence of an over-controlling central authority. It requires a certain amount of distributed control to allow other principles of CAS to occur.

vi) **Shallow Structure**: components of a CAS should be organized and arranged with the least amount of structure necessary for the system to efficiently achieve its objectives.

vii) **Growth & Evolution***:* the CAS and its components respond to changes in both the internal and external environment. Adaptation occurs through continued growth and evolution in small increments. This adaptation allows the CAS to pay attention to the needs of its immediate surroundings and in doing so re-orientate itself (with incremental steps).

2.2.2. Untangling the Mechanisms of IS using CAS

The third-wave of systems theory considers CAS to be highly analytical and superior as it can be used as an instrument for "untangling" the mechanisms and processes of IS (Nan, 2011). CAS theory is commonly used to describe environments of agents, components, interactions and reactions in relation to one another so that emergent system behaviours can be demonstrated. Schneider and Somers (2006) provide hints that emergent behaviours and patterns are a result of "strange attractors".

CAS theory can be used as a lens to examine strategy development processes (Hammer et al, 2012), or even as a tool to model system components in IT use-processes (Nan, 2011). Sturmberg et al. (2012) demonstrate the needs and relations of agents across different levels of a CAS in relation to a central *attractor* or vision for the purpose of conducting a policy analysis.

Complex adaptive systems are difficult to predict and it is recommended that descriptions across multiple scales be developed to understand underlying mechanisms and dynamics (Merali et al, 2012). Hammer et al. (2012) suggest that issues of complexity are, for the most part, irreducible and the practical value of CAS theory for organizations is to understand how to live with complexity rather than to reduce it. Typically it is recommended that empirical studies of systems be used to explain dynamics and introduce alternative strategies of intervention where necessary (Mills et al, 2003). Daft (1992) in Anderson (1999) equates organizational complexity with sub-systems and activities but notes that complexity can be measured across three dimensions, namely vertical (the number of organizational hierarchy levels), horizontal (the number of departments or job functions per vertical level) and spatial (geographic locations). In large organizations and enterprises adopting a CAS view of internal systems could help provide insight and understanding into processes and practices.

Meadows (1999) introduces us to leverage points (within complex systems) which can be described as those places where small changes or shifts can have major influences over the entire system. They are generally considered points of power and may exist as policies or even as relationships between people. They tend to be known by many but are often misunderstood. As a result they are frequently pushed in the wrong direction. True leverage points may at first appear counter-intuitive creating misbelief about their positions but they can be learned by studying a system (Meadows, 1999).

Hammer et al. (2012) define and use CAS as a lens to examine strategy development processes within organizations. Hammer et al. (2012) describe each of these four facets as follows: *continuous varying interactions* (CVI), *patterns development* (PD), *people factors* (PF), and *self-organization* (SO) which occur throughout a CAS as ever changing phenomena (see Figure 14).

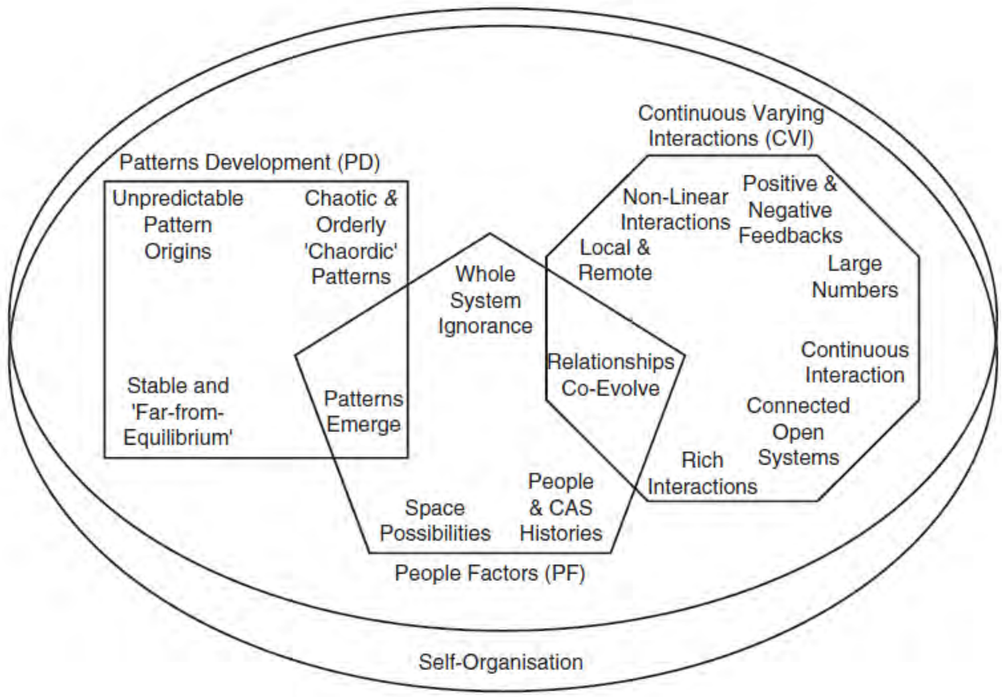


Figure 14. Four facets and 16 characteristics of a CAS lens used to examine strategy development processes (Hammer, Edwards & Tapinos, 2012).

Characteristics of *continuous varying interactions* (CVI) include:

1. Local & Remote (interactions): the local relationship network of the organization (or system) is where the richest interactions occur with influences having far reaching effects; (remote) connections or relationships are considered very important due to their non-linearity.
2. Non-Linear Interactions: (system) interactions are unpredictable with regards to cause and effect relationships; small actions can result in big effects and visa-versa; the scale of actions and interactions are totally unpredictable.
3. Positive & Negative Feedbacks: both restraining (negative) and developmental (positive) feedbacks can be present.
4. Large Numbers: many different elements (people) and subsequent relationships (between people) are present.
5. Continuous Interactions: there are endless, repeating and dynamic interactions between people within and external to the system/organization.
6. Connected Open Systems: a CAS is an open system and can interact actively or passively with other CASs; these interactions can occur across various levels (of integration).

7. Rich Interactions: quality of interactions is varied, always changing, iterative and self-referential.
8. Relationships co-evolve: as people and the CAS develop - a variety of relationship “rules” will co-evolve influenced by traditions, customs or organisational “culture”.

These characteristics offer different perspectives on the types of relationships that exist between people or groups of people within a CAS. They describe dynamic connections that refer to CAS “structure” but the connections are never static and are constantly evolving.

Characteristics of *patterns development* (PD) include:

1. Patterns Emerge: coherent patterns of “order” emerge spontaneously becoming *attractors* that may further influence the development of the pattern and the CAS.
2. Stable and far-from-equilibrium: a CAS will adapt and survive through periods of turbulence, whether internal or external. It is noted that stability is not a prerequisite for progress and can lead to atrophy.
3. Origins of patterns: patterns are unpredictable in time and place - they are spontaneous.
4. Patterns and Attractors: can be orderly (stabilising), chaotic (de-stabilising) or both simultaneously (‘chaordic’).

These characteristics offer perspectives on whole-system behaviour. It seems some level of instability is required to inspire the development of new patterns and attractors for a CAS. Without this patterns development would not succeed.

Characteristics of *people factors* (PF) include:

1. Whole system ignorance: a single person is incapable of having complete knowledge of the CAS due to its complexity and dynamics. Therein lays the risk of uncertainty that affect people and organizations of the CAS.
2. Histories: the origins and histories of development are very important, especially so for people and the CAS. They influence development options, choices and future actions (referred to as “Path Dependency”).
3. Space possibilities: CASs are embodied by people that develop by adapting to existing conditions. They are capable of exploring “space” (and time) by thinking, learning, imagining and making decisions.

These characteristics offer perspectives on the complexity of people, their histories, limitations, and capabilities. They influence actions, decisions and outcomes.

The fourth facet “self-organization” underlies the others. It is ever present and always shifting in importance because it is influenced by internal and external factors that are continually changing. In addition to these factors different management control approaches (e.g. loose or tight) may be practiced by different people and managers (Hammer et al, 2012).

2.2.3 Systems Thinking and Public Health

Public health dynamics have been a focus and concern to leaders for over a century now and systems thinking has provided much insight into understanding and identifying its afflictions (Trochim, Cabrera, Milstein, Gallagher & Leischow, 2006). Our ability to work towards affecting change in this field has improved through innovation in concepts, methods and moral frameworks. These new practices are affecting the public health discipline in ways that the early pioneers could scarcely have imagined.

Systems thinking is not a single discipline, but should be thought of as a series of interlinked disciplines (Trochim et al, 2006). It emphasises a focus on the relations between different structures of a system and the different levels and scales of its interrelated structures. Examples include our human bodies, our home environments, our work environments and even our political systems with their governing structures. It relies on a common conceptual paradigm that considers connections between different components, plans for the implications of their interactions, and requires transdisciplinary thinking and active engagement of those that have an interest in the outcome, to govern the course of change.

It is argued that complexity in public health poses a formidable challenge and the use of studies grounded in explicit systems orientation, such as systems thinking and modelling, may provide the type of innovative responses necessary to overcome these challenges (Trochim et al, 2006). By learning how public health exists as a system of structured relationships, how it is organized, how it behaves over time and how it can be better governed, we may be able to analyse and synthesize a response to current obstacles in the field.

Sturmberg et al. (2012) demonstrate and visualize a policy analysis of the Australian healthcare system using CAS (see Figure 15) based on the metaphor illustrated by Capra (1996) of the bathtub vortex. The system *attractor* in this metaphor (a `good personal health experience` by the patient) would be the plug hole of a bathtub into which water flows in the form of a vortex. Within this vortex different organizational levels exist, each with different levels of complexity, interaction, and certainty/predictability (i.e. macro, meso, micro, and nano levels). Complexity and interactions increase exponentially towards the vortex attractor while certainty/predictability decreases. Further away from the vortex attractor these attributes become less complex and more predictable. Sturmberg et al. (2012) point out that a disturbance to the vortex will always result in the restoration of the vortex close to its original form/position.

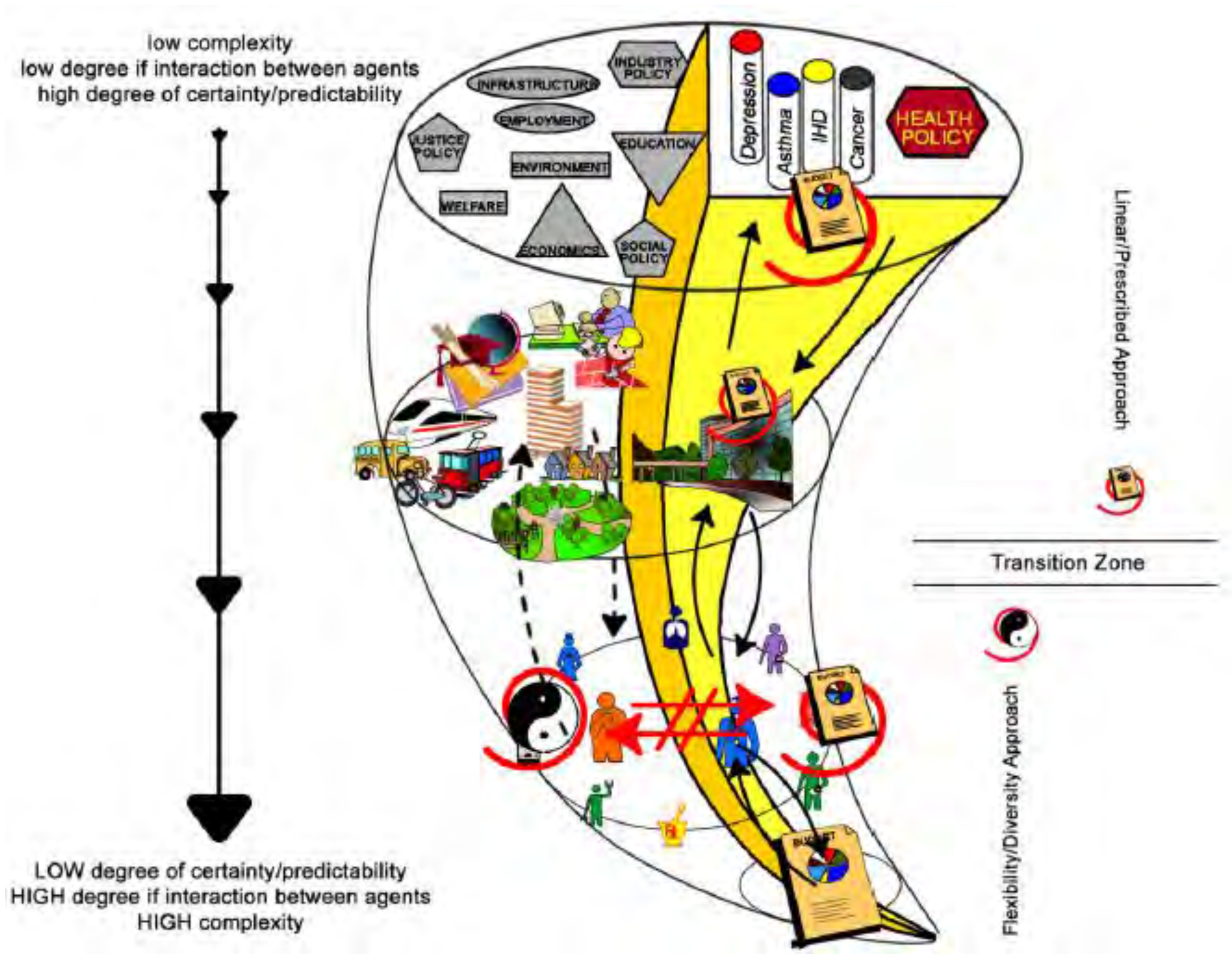


Figure 15. The ‘healthcare vortex’ of Australia’s healthcare system that is driven by budget and disease-specific concerns (Sturmberg, O'Halloran & Martin, 2012).

2.3 Health Information Systems Domain

As the case study for this research project takes place in the public health information systems domain, some background to health information systems (HIS) is presented. This includes a review of the field known as health management information systems (HMIS) and a brief review into routine health information systems (RHIS). The concept of a public health observatory (a first world country institution) is also presented.

2.3.1 Health Information Systems

The goal of any country level health system is to improve the health status of its people (Lippeveld, 2001). Conceptually Health Information Systems (HIS) is positioned amongst information systems that process data, information and knowledge in health care environments with the aim of contributing to high-quality and efficient patient care (Haux, 2006). They were originally focused on collecting information about diseases (“surveillance”) and health service outputs (Lippeveld, 2000). The term “health management information systems” can be misleading impressing the idea that different types of information systems exist with different functions (e.g. management, epidemiological, administrative, etc). These individual information system ‘types’ are considered sub-systems of an integrated HMIS because they generate information that can be used to improve health care management decisions across all levels of a health system (Lippeveld, 2000).

A well-functioning HIS will support functions relating to i) individual (patient) care management, ii) health-unit management, and iii) health system management (Lippeveld, 2001), where the ultimate objective is to “improve action” rather than simply “collect information”. This is done by monitoring and evaluating processes to ensure adequate inputs are present so that they produce desirable and timely outputs. Different information needs exist which are identified and adapted over time as part of planning and management processes. This affects data collection, transmission, processing and analysis components that are part of the ‘information process’ (Lippeveld, 2001), or, using a more up-to-date version of that same model, the ‘data handling processes’ (Aqil, Lippeveld & Hozumi, 2009). Process structure and design is critical for each component of the ‘data handling process’ (see Figure 16) as a weakness within any one of these components will reduce the effectiveness and relevance of information within the HIS (Lippeveld, 2001).

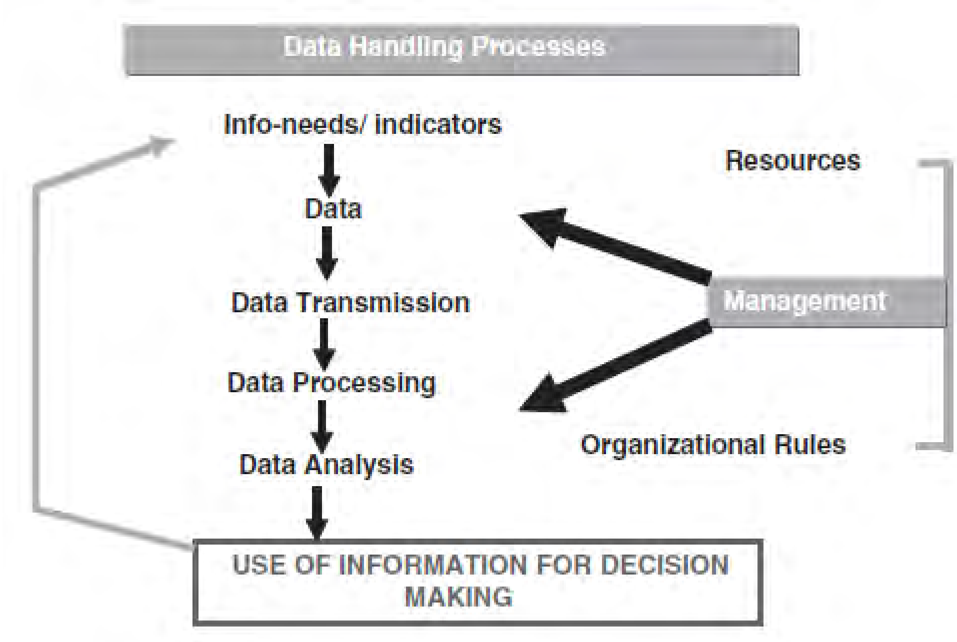


Figure 16. Health Information System components diagram (Aqil, Lippeveld & Hozumi, 2009).

Lippeveld (2001) classifies health information systems according to data collection methods: routine and non-routine in which routine health information systems (RHIS) collect patient/client encounter data. This data can be collected at health facilities, through outreach services or within communities. Non-routine systems collect data through surveys, using quantitative and qualitative rapid assessment tools or through specialised studies.

In the public sector HIS are used by practitioners of public health which can be described as “collective action for sustained population-wide health improvement” in which attempts are made to measure and monitor the health of a country’s population (AbouZahr & Boerma, 2005, p. 578). Information access and use is at the core of this discipline, and practitioners in the field of Public Health are heavily reliant on data so as to understand and interpret the changing patterns, with the aim of adapting and positioning health services where required.

One crucial goal of any health information system is to influence policy making, programme action and research by collecting, processing and reporting on health information and knowledge (AbouZahr & Boerma, 2005). However, there is much that remains unknown about adult mortality, causes of death and other related health outcomes. AbouZahr and Boerma (2005) present a set of domains that any health and information system should address (see Figure 17).

Figure 17. Various aspects health information systems should address (AbouZahr & Boerma, 2005)

Health care organizations are complex environments in which numerous functions, processes and roles converge, where objectives diverge and where leadership and power sharing may be dispersed (Mills, Rorty & Werhane, 2003). The 'system' of health care organizations may implement processes and practices that are designed to produce predictable results. There may be strong consensus on what outcomes are desirable but these systems are not mechanical in nature and are comprised of human beings with the freedom and ability to respond to stimuli in different and unpredictable ways. As a result the outcomes from these systems tend to be less predictable.

Health systems are under continuous pressure to improve quality (particularly their processes and outcomes) while working with constrained financial resources (Mills et al, 2003) and new management approaches have been called for. The Centre for Disease Control (CDC) made a statement signalling the need for health system workers to adopt a systems thinking and modelling approach while crossing their traditional boundaries in the effort to manage complex challenges (Leischow & Milstein, 2006**)**. Complex adaptive systems theory seems an obvious choice for exploring this environment.

2.3.2 The Public Health Observatory: a Health Intelligence Institution

In 1974, in France, the first public health observatory was established to support the field of public health and social care (Hemmings & Wilkinson, 2003). This 'institution' was tasked with supplying regional information for health-policy decision making. In addition to this, officials were able to engage with the institution and make enquiries to identify topics of concern. They were also tasked with the development and execution of projects. Later in 1990, in Liverpool, the first English public health observatory was established. This was modelled on the French public health observatory but went further than only providing information. It also provided contextual perspective on its findings to ensure relevance and accuracy. There has been much deliberation on the use of the word "observatory" but it was chosen because the role of the institution was to "stand back from phenomena and events, providing objective description and analysis, and forecasting of patterns, interrelationships, processes and outcomes" (Hemmings & Wilkinson, 2003).

For the most part these 'institutions' are difficult to define as they are products of their historical and contextual circumstances (Hemmings & Wilkinson, 2003). Organizationally, they may be small and independent with strong academic qualities that are merged with state-based public health. They are able to determine, collect, integrate and synthesise data, yet do not necessarily collect and store data in repositories (see Table 5). They provide high quality information in short timescales and have networks that are able to access different data sources. As a result of this networking ability, they can be considered greater than the sum of their parts, a fact consistent with systems thinking. In summary, they assist people to identify and deal with different health related issues, which is crucial in supporting policy making.

Role	Example
Monitoring health and disease trends and highlighting areas for action	Working together on coronary heart disease - a focus on the inequalities existing in coronary heart disease, together with recommendations for action
Identifying gaps in health information	Perinatal and infant health: a scoping study to identify current information sources and the gaps that exist
Advising on methods for health and health inequality impact assessment	An overview of health impact assessment
Drawing together information from different sources in new ways to improve health	Towards a healthier region - this health profiles uses housing and employment data alongside health data
Carrying out projects to highlight particular health issues	The dental health of five-year-olds
Evaluating progress by local agencies in improving health and cutting inequalities	Baselines have been established in projects/reports and others on key issues. Trend data will be published in future
Looking ahead to give early warning of future public health problems	Future conference - a forum for partners to address likely future public health issues such as aging populations and genetics

Table 5. The roles of public health observatories in England with examples (Hemmings & Wilkinson, 2003).

2.4 Literature Summary

Intelligence is the measure of information processing required to make decisions that achieve goals. Information processing requires data generation according to informational-relationships that support integration. Integration leads to data transformation which is typically made accessible in data warehousing environments. From here data is made accessible and applied to relevant situations. Different analysis processes exist along with a variety of implementation architectures but four major focus areas have been drawn out: information quality, master data management, warehousing architecture and analytics. These areas of BI have been discussed in detail and are important for the direction of this research project.

Complex adaptive systems theory focuses on people and groups, their relationships, abilities to self-organize and general characteristics that facilitate learning and adaptation. The theory is considered highly analytical and has been shown used to enquire, model and demonstrate emergent behaviour. It has also been shown as relevant to the area of health systems and by extension health information systems. Most importantly it can be used to untangle mechanisms of IS. Health information systems are used to support policy making, programme action and research. Different types of HIS exist with different purposes. Routine health information systems collect aggregated data on a scheduled/routine basis for the purpose of collecting patient/client encounter data. Health information system components have been presented along with various aspects that health and information systems should address. The public health observatory, a health intelligence institution, has also been mentioned.

3. Research Design and Methodology

The following is a summary of the research methods used in this study. Exploratory research is presented which is followed by a review of interpretive methods. This is followed by a discussion on qualitative reasoning and case study research. The final section summarises the research process as it unfolded and include a description of the data analysis process describing contributions by the different participants.

3.1 Exploratory Investigation for Discovery

In order to achieve the objectives the environment needs to be explored, i.e. an **exploratory** *investigation* will be done for *discovery*. Stebbins (2001) defines exploration for research in four different categories (see Table 6). The type of exploration to be employed by the researcher is a combination of 'investigative exploration' and 'exploration for discovery' as the area of business intelligence for public health is not sufficiently documented in academia. In order to exemplify BI phenomena in the case study the researcher is required to conduct a broad investigation of the interactions, agents, processes and resources being exchanged. The types of interactions, feedback loops and emerging behaviours of the CAS's will be explored for the purpose of discovery. The research process will be repeated until nothing more can be found or discovered about the changing processes as they relate to current thinking around business intelligence.

	TYPE	PURPOSE	REQUIREMENTS	
1	Investigative Exploration	to understand	Inquisitive examining	
2	Innovative Exploration	to create output effect or product	Gaining familiarity with substances & procedures needed to manipulate them to achieve desired effects	
3	Exploration for Discovery	to discover or experience	Persistence - broad and thorough investigation repeated until nothing more (that is required for understanding or describing) can be found or discovered	
4	Limited Exploration	to satisfy explorers interest in searching for something particular	Clear understanding of desired outcomes	

Table 6. Four types of exploration (Stebbins, 2001).

3.2 Utilization of Interpretive Methods

"Interpretive research can help IS researchers to understand human thought and action in social and organizational contexts; it has the potential to produce deep insights into information systems phenomena including the management of information systems and information systems development." (Klein & Meyers, 1999, p. 67). Walsham (2006) discusses interpretive methods and considers them from the point of view of the researcher. He describes these methods, suggesting that our theories of reality are our way of interpreting and understanding the world. Within these theories we may find shared meaning or similarities in different subject material and for an interpretive researcher these similarities are based on reality, not our own interpretations thereof. Research qualifies as interpretive when knowledge is gained only through social constructions such as language, consciousness, shared meaning, documentation, tools and artefacts (Klein & Meyers, 1999). It focuses on complexity of human sense making as situations emerge. Walsham (2006) describes interpretive research in the following activities:

i. Carrying out fieldwork

This aspect forms the basis of all interpretive research and comprises of sub-activities such as choosing a style of involvement, gaining and maintaining access, collecting field data and working in different contexts. Walsham (2006) recommends the researcher be present in body and spirit; demonstrate sincerity, honesty; demonstrate respect to achieve a mutual trust. The researcher should embrace these attributes to the best of their ability.

ii. Theory and data analysis

Theory selection provides an initial guide for the design and collection of data; this can be done as part of an iterative process for data collection and analysis; or it can be applied as part of the final phase of the research. For this project, the researcher will be relying on the theory of Complex Adaptive Systems (CAS) to identify and interpret changing processes. CAS and systems thinking will guide the data analysis process by suggesting structures upon which to model the emerging phenomena. The following tools or processes can be used for the data analysis and interpretation:

- Logical Analysis, Analytic Induction and Hermeneutical Analysis
- Systems Thinking influenced heavily by Complex Adaptive Systems Theory

iii. Constructing and justifying the contribution

This area of work is applied post-fieldwork and consists of activities related to the communication of one's work. Primarily its main focus is on the construction and writing of the dissertation for the purpose of gaining recognition by others. The process can be regarded as justifying the approach, constructing the contribution and writing.

iv. Ethical issues and tensions

Although very little is written on this issue Walsham (2006) describes some of his concerns: harm to participants, lack of consent, invasion of privacy, and deception. Fully addressing these concerns in practice can be difficult to achieve but Walsham provides many examples of how this can be managed. Examples include:

- making participants identifiable: although sometimes impossible to avoid (due to informed guess work) try to avoid identifying people by name or position without diminishing contextual information;
- power relations: when working with the organisation avoid reporting on things they do not want reported;

In addition to these recommendations the researcher should, as far as possible, adhere to the rules laid out by the UCT ethics committee.

3.3 Qualitative Research

Hyde (2000) discusses reasoning when conducting **qualitative** research. He draws a comparison between inductive and deductive reasoning by describing the inductive approach as a process in which one seeks to observe instances or patterns and draw out generalizations or theories. Deductive reasoning is a theory testing process which either confirms or discards the proposition.

Cross-sectional research is used to describe a group of subjects at the same point in time and may be used to draw generalizations (Campbell, Machin & Walters, 2010) while longitudinal research emphasises the study of change (Ployhart & Vandenberg, 2010). Longitudinal research requires a minimum of three repeated observations on at least once of the substantive constructs of interest. "If a process of change is an important aspect of what is being researched, and especially if the processes involved are complex or the timespan substantial, a single episode of data collection may not be enough. For this type of study qualitative longitudinal research may be more effective" (Ritchie, Lewis, Nicholls & Ormston, 2013).

3.4 Case Study Research

"Case study research is one of the principal means by which inquiry is conducted in the social sciences." (Thomas, 2011, p. 511). Its definition varies according to practitioner background but those from medicine and law regard a case study as a vehicle for demonstrating or illustrating novel phenomena. Yin (2002) discusses the need for case study research and observes its use out of the desire to understand complex social phenomena. It allows investigators to observe and retain meaningful characteristics of events such as organization and managerial processes, individual life cycles, system changes, system relations and the evolution of industries.
"Case studies are analyses of persons, events, decisions, periods, projects, policies, institutions, or other systems that are studied holistically by one or more methods. The case that is the subject of the inquiry will be an instance of a class of phenomena that provides an analytical frame - an object - within which the study is conducted and which the case illuminates and explicates." (Thomas, 2011, p. 513)

Yin (2002) provides guidelines for case study research:

- it is an empirical enquiry into contemporary phenomena within real-life contexts where boundaries between phenomenon and context are not clearly evident;
- it relies on multiple sources of evidence with data needing to converge in a triangulating fashion;

Case study data can be collected from the following sources: documentation, archival records, interviews, observations, participant-observations and physical artefacts. Used together multiple sources can be used as evidence to provide validity and reliability.

3.5 Summary of Research Process

This research project looked towards the evolving processes within a complex environment over a time span exceeding 15 years. The substantive constructs were the four focus areas of BI identified by Tan et al. (2011) in Table 1. For this reason a longitudinal approach to data collection and analysis was chosen.

There was no need for theory testing therefore an inductive approach was chosen. An exploratory investigation looked to find shared meaning or similarities in different subject materials, namely business intelligence and the area of health information systems. The researcher's style of involvement varied from 'outside researcher' (focusing on formal and informal interviews with case study participants) to 'involved researcher' as the researcher's perspective included details from the case study dating back to 2001. There were phases of both formal and informal interviews as well as informal feedback discussions. The majority of Interviews were with colleagues from the Health Information Systems Programme (HISP). Two permanent employees from the National Department of Health participated, with one full-time consultant giving time and inputs.

Gaining and maintaining access was supported by the researcher's personal involvement in the case study project. The researcher did not pursue an action research approach but is aware that the research process has provided new insights for future areas of development.

Field data was collected through documentation, interviews, observations as well as physical artefacts. Project documentation was obtained with permission from HISP; a series of interviews were conducted with the numerous participants (see Table 7); observations through informal discussion assisted with research direction and data collection; physical artefacts were known to the researcher which formalize and structure the later part of the case study environment. For an example of interview questions see Annexure H.

Participant	Role	Background
1 (researcher)	BI Developer (software & database)	IT, IS, Researcher
2	Data Analyst/Scientist, GIS Specialist	GIS, IS, Academic
3	Developer & Infrastructure Manager	IT
4	Data Analyst, Public Health Specialist	Public Health
5	Executive Health Administrator	Public Health, Administration
6	Data Analyst	Public Health
7	Health Administrator	IS, Public Health
8	Executive, Health Specialist	IS, Public Health, Administration, Academic
9	Public Health Specialist	Public Health
10	Public Health Specialist	Public Health

11	IT Contractor	IT
12	GIS Specialist	IT, GIS, Administration
13	Public Health Specialist	Public Health
14	Public Health Specialist	Public Health, Administration

Table 7. The Fourteen participants arranged according to role and background.

Initial interviews began with a standardized set of questions that were adapted according to background, history, knowledge of the environment and experience in the case study. This first round of interviews was more focussed on discussions around BI processes as specialized areas of information management but they lacked depth which resulted in poor data. The research thematic was not well structured yet and this required a deeper investigation into BI themes in literature.

Through a series of informal discussions (of trial and error) the four focus areas identified by Tan et al. (2011) emerged as coherent and suitable themes that provided grounding for the case study's history (see Table 1). They were consistent with interviewee dialog, historic challenges and current developments recognized by the researcher (who participated in the development of the RHIS and SSBI tool). This realisation was significant for the analysis process as it provided clearly defined themes for the interpretation of events according to timelines. Because of the longitudinal nature of the case study different participants were able to provide insights into these focus areas. For example participant 14 (see Table 7) had tremendous experience and interest in data quality issues. This provided the spark for the DQ theme from which many other interviews provided data (e.g. participants 10 and 13's inputs). Participant 2 was present throughout the lifetime of the RHIS system from its prototype conception in 1998 up until today (2014). His feedback together with literature and inputs from participant 8 (who contributed as a referenced source) provided the spark for the MDM theme around which follow up discussions would take place. Participant 2 also provided inputs into the DW and Analytics background as well as important insights into the historic setting of the health system which predated the establishment of the RHIS. Participants 11 and 12 were responsible for the web reporting system concept that emerged as the SSBI tool from literature (see Figure 4). Their team-work, discussions and ideas were responsible for its eventual development which emerged from interviews. Only after a deeper review into BI would the SSBI concept be considered as a theme for this study. Participant 7 provided insights into the use of information within the DOH which supported the development of the decision making component from Annexure I (which supplemented inputs from participant 13 on the presence of resource allocation authorities in developing countries). Other participants provided valuable insights and historic information that supported these emerging themes. In summary no thematic analysis was conducted ahead of first round interviews. Only after a deeper, second round literature review was completed did informal discussions take place that helped establish a theme structure.

Informal discussions began in late 2011 and early 2012 which coincided with the NHIRD implementation phase. This was required to determine the research path in terms of IS focus areas. Over a fifteen month period a total of 20 voice-recorded (formal) interviews were conducted which were followed up by at least 10 informal discussions (non-recorded). The majority of interviews took place in person with some being conducted over Skype or telephone. In total 13 people participated in 20 formal interviews (see Table 7) providing a total of 947 voice recorded minutes. Formal interviews began in May 2013 with the final concluding interviews taking place at the beginning of July 2014. Issues of privacy regarding information disclosure were raised throughout the interview process. At no time throughout these interviews were concerns of a personal, organizational or political risk raised by the participants. In addition the researcher adhered, as much as possible, to the rules laid out by the UCT research and ethics committee. No ethical issues or tensions were raised throughout the research period. The majority of participants provided input into evolving processes and were not required to disclose information of a sensitive or personal nature.

4. Findings

The following sections are presented according to the major focus areas of BI as identified by Tan et al. (2011). Each section is presented as it evolved in a longitudinal format within the respective subject area of BI as opposed to a combined overview as that would distract away from repeated observations within each area of interest. Data warehousing and analytics are presented together as they are heavily co-dependent in the case study.

An introductory section will describe this researcher's personal involvement in the development of the routine health information system and the eventual self-service BI tool. This is followed by an interpretive description of the environment as modelled in the data warehousing environment (with aspects of the data model development). These interpretations are based on the presence of structures known to the researcher. This lays the foundation for other activities as it has supported the development of data collection, integration and analytics capabilities in the case study. The evolution of master data management is presented followed by the evolution of data quality. The final chapter is a review of data warehousing and analytics capabilities as a combined section.

Findings are based on a combination of documentary (secondary) data and interview (primary data). This was done out of necessity to help support a process of triangulation that was made possible with the researcher's personal account of experiences in the case study.

4.1 The Case Study: My personal experience and point-of-view

My involvement with the DHIS 'project' began during my "early years" in IT (1999 - 2004) in a software development role. In 1999 I entered the IT arena as a junior programmer and began working as a Visual Basic developer for a small IT firm in Cape Town. Around 2000 I became involved in this project. This prototype application, known as the District Health Information Software (DHIS), was a 'stand-alone' distributed database application that collected aggregated (routine) data from different health facilities to form overviews at regional levels. I was immediately attracted to this prototype and its development process. Reflecting back now I can surmise that the idea of "shared knowledge and distributed ownership" was the attractor. The vision and philosophy embodied by the people managing the system development process in terms of content and design was inclusive. As a junior developer my ideas and thoughts were relevant and sometimes incorporated into the software. Rather than being governed by an authoritarian styled control system (with rules and relationships determined by complex power structures) it was dominated by an inclusive bottom-up approach that nurtured participation through questioning, discussion and knowledge sharing. This development style attracted an honesty which rallied around a vision to strengthen the South African health system. Ownership was distributed which was a relief from the 'private' contract work experience at some of the larger organizations where freedom to participate was governed by unwritten rules that were mostly unknown.

The DHIS version (1.3) was a relational database system with a flexible data model. Any conceivable health phenomena at an aggregation-level, i.e. counts of health system encounters (see Annexure A), could be defined as a data element and could be recorded at any facility found in the database collection of 'organisation units'. This facility list was structured according to South Africa's administrative hierarchy, i.e. province, district, sub-district and health facility. Data elements were developed in a consultative process by public health and health programme specialists and were arranged according to health management or health programme concerns. The data model supported the design of aggregation-rules that were defined as indicators with numerator and denominator formulas (see Annexure B). Once data has been entered a transformation process could be run to export calculated data into a separate database known as the data mart (which signified the beginning of the ETL and data warehousing concept within the system). This aggregated data was then fed into Excel pivot tables and became the process through which data was made available and shared (i.e. early spread-marts). Throughout my time at this firm, the unfolding developments of DHIS were shared back into other software products and visa-versa. The DHIS v1.3 grew to include geographic information system (GIS) integration capabilities allowing aggregated data to be exported and viewed in a version of Arcview GIS. Around the same time an early version of a data dictionary (web-application) was developed to share data element definition management. Unfortunately this component may have been developed ahead of its time as data-governance processes in 'health' were still developing. The DHIS software version 1.3 was under permanent development and redesign between 1998 and 2003. During that time the software was prototyped and incremental developments gave rise to a wide range of functional components as demonstrated in their software introduction slide (see Figure 18).

Based on Microsoft Office Professional

Uses Access for Data Entry and Excel Pivot Tables for data display,manipulation & graphing

Contains

- Facility lists (with institutional infrastructure)
- Data dictionary
- Data parameters
    - Max/min
    - Validation rules
- Indicators
- Community data (population)

Figure 18. DHIS version 1.3 components from the early 2000's reflective of a decision-support system.

By 2004 Health Information Systems Programme (HISP) was legally established as an NGO (a not-for-profit organization) and they recognized the need for a full time software/database developer. Around mid-2004 I began work on the new, more flexible version of 1.3. The prototype had proven itself by developing a proven set of processes, concepts and content. It had a significant 'foot-print' in the health-management domain but with the ever growing requirements from its broad user base it had reached the limits of the data model. This signalled the beginning of a new design phase and a new iteration of the software (version 1.4). During June of 2004 a workshop was held in Cape Town by my future 'supervisor' (a scholar, data scientist and GIS specialist) in collaboration with the University of Oslo. Several IS masters students from Oslo University gathered in Cape Town in a little house in Rondebosch to plan the way forward. During the workshop two alternative streams for the software were proposed: a desktop MS Access database application DHIS version 1.4 (DHIS14), as a continuation of DHIS version 1.3, and a web-based, open-database platform which would be known as DHIS version 2 (DHIS2) which would be built in Java.

After a week-long workshop of intense planning, arguing and negotiating, a single database design emerged that was relational and would be the cast upon which these two systems would be developed. The University of Oslo has a long and proud history with the philosophy of open-source and it was a prerequisite for all involved. Over the next two years we would work, rework and adapt the database model and its software making tweaks and redesigns to meet the needs of users already familiar and happy with the previous version's functionality.

By 2006 we had rebuilt the data management and maintenance sections, data entry was stabilizing and we had completed the redesign of the ETL process (familiar to everyone as the 'export to data mart' process). Due to the flexibility of the system design outputs to the data mart were sometimes unpredictable but this was mostly due to the dynamic modelling of organisation unit dimensions (an unanticipated consequence). Soon, standardized data rules were applied to limit the classification of organization unit super-types into 'group sets'. We began work on the automation of pivot-table outputs based on standardized queries within our data marts. Our NGO had grown from 5 to 11 staff and our workload was increasing. New projects were leading to new requirements and functionality.

South Africa has been fortunate to have so many development partners participating in health improvement projects. Many of these partners (in health) have participated through the implementation and development of ICT. Others have provided health services, professional services and in some cases drug stock supplies. One development partner (known as the Italian Cooperation) contributed hardware, software and professional services. Between 2004 and 2005 they supported the development of ICT tools in the Gauteng provincial department of health (GPDOH). During their time there they proposed the development of a web-based reporting module based on the DHIS data mart design. Data was being provided regularly through spread marts and the project concept was supported by the GPDOH.

Around 2006/2007 HISP suffered through a financial crisis and were required to take on a range of projects to sustain themselves through the 'lean' times. GPDOH had obtained funding for their web-based reporting tool and requested a meeting with our NGO after which we agreed to begin work on the web-reporting concept. Although clear deliverables were negotiated (e.g. reporting outputs, graphics capabilities, GIS integration, color-coding for different indicators, etc.) many of the system components were designed and adapted on the job. The end product was a web-based reporting and graphing interface running against a SQL backend that supplied data from the regular data mart. Because the data mart was reloaded on a monthly basis, the SQL database would be reloaded monthly following the spread-mart distribution cycle. By this time the Italian Cooperation were focusing efforts in the Eastern Cape Department of Health (ECDOH). Hearing of the successful development of the reporting module at GPDOH through a development-partner network, the Italian Cooperation began efforts to implement a roll-out at ECDOH. Unfortunately HR capacity constraints hindered the rollout resulting in a temporary implementation of 6 months. Although the implementation could not be sustained there were successful developments and learnings:

- the implementation would require dedicated resources to ensure data access on a technical level;
- the implementation would not change ‘information-use’ culture in such a short time;
- feedback mechanisms are critical for defining user-needs;

In 2010 the NDOH made informal requests to different provincial offices. They were interested in looking at what progress had been made with regards to ICT tools to assist users with data interpretation. These events led to the presentation of the GPDOH web-reporting solution to executive level managers from the NDOH. “It wasn’t a formal thing. It was more like an informal update on progress”.

Unknown to everyone, executive management at NDOH had put forward a request to establish a centralized data warehouse to serve health managers. “The whole data warehouse idea came from the DG. She drove it. That was her vision”. This single presentation led to a meeting at Parliament offices in Cape Town in which HISP were drawn into a strategy for establishing a centralized warehouse for health. Two months later a proposal was drafted and signed giving the go ahead for the establishment of a project which would later become known as the NHIRD (National Health Information Repository and Data warehouse). The NHIRD project started off in early 2011 with members of the project-team taking part in various training ‘exposure’ sessions.

4.2 Modelling an Evolving and Complex Environment

The South African public health system is a vast, complex environment comprised of various entities, organizations and groups of people that work together to provide health services to the public. CAS theory teaches us that systems are greater than the sum of their parts (Shaw, 2009) and that it is impossible for one person to have complete knowledge of all system components and dynamics at any one time (Hammer et al, 2012). These components and dynamics are too numerous to mention but the following groups have been drawn from the case study:

1. International donors and funders, e.g. aid agencies that provide financial inputs and are positioned as resource management and allocation authorities;
2. Local donors and funders, e.g. private sector sponsors and government sector agencies such as Treasury that provide financial inputs and are positioned as resource management and allocation authorities;

3. In country health service administrations, e.g. health ministry institutions that coordinate health services through their administrative structures;
4. Development and implementation partners, e.g. NGO's, NPO's and other development organizations that support the development and implementation of health services and administrations;
5. Public health care services, e.g. clinics, hospitals, associated health services and staff;
6. Populations, e.g. persons, patients or populations utilizing health care services.

It is possible to describe complexity using a sub-systems approach (Anderson, 1999). Sub-systems can be demonstrated across three common dimensions namely *vertical*, *horizontal* and *spatial*. In this case study the routine health information system has, since the late 1990's, relied on standard dimensions which have evolved and matured. Together they have helped in the development of a standardized data model within data collection and warehousing environments.

4.2.1 The Vertical Dimension

The public health system of South Africa can be separated into different vertical levels representing a mixture of "public health service administrations" and "public health care services". This mix of administrative "regions" and health service levels are known in the data warehousing environment as "organisation unit levels". Various implementations of these structures have been developed for different data sets but the most common set modelled and defined in the routine health information system (DHIS) follow the :

Level 1: national (country);

Level 2: province;

Level 3: district;

Level 4: sub-district (sometimes known as local-municipality);

Level 5: health facility;

Below level 5 lie a variety of organizational unit levels that vary according to the information collection systems or data sets under representation, e.g. in some cases outreach-teams and team members are represented as level 6 and 7 respectively while in other cases a hospital-ward could exist at level 6.

Different information collection systems, programs or initiatives reserve the names of these vertical levels, adding variety to data warehousing needs and introducing new challenges for data managers (see Annexure C).

The vertical dimension is very much aligned with the spatial dimension as defined by the demarcation board of South Africa (with the exclusion of the local "political ward" level). The administrative system in health follows the spatial dimension neatly. The absent "political ward" level is currently under review for inclusion into new data sets and structures, introducing a new challenge for data managers.

4.2.2 Horizontal Dimension

A variety of *organisation units* exist across the different vertical dimensions. At present (2014) level one to four are standardized according to the RSA "Health" administration hierarchy (i.e. provinces, districts and sub-districts). Level five (health facility) has the highest number of organisation units and the greatest variety. The health administration has 2 major classifications for these health service level organisation units:

- *Primary Health Care (PHC)* organisational units provide the widest range of health services utilizing nursing staff and other health professionals but not doctors. These health services are usually the first line of support for persons seeking medical help under the national health system. They are usually small in scale (only employing a few staff) and are centrally managed by their sub-district, district or provincial level offices. Examples include:
    - Clinics (including satellite)
    - Community Health and Community Day Centres
    - Correctional Centres (supporting prisons)
    - EMS Stations
    - Frail Care and Hospices
    - Industry Clinics
    - Mental Health Services
    - Mobile services

- Hospital services are organisational units that provide a variety of specialised health services. They employ a wide range of health professionals including nursing staff and medical doctors receive referrals from PHC services. Often these organisational units are extremely large (e.g. Chris Hani Baragwanath Hospital in Gauteng is the third largest hospital in the world) and require their own administrative structures to manage services. Examples include:

- District Hospitals
- Military Hospitals
- National Central Hospitals
- Provincial Tertiary Hospitals
- Regional Hospitals
- Specialised Hospitals

4.2.3 Spatial Dimension

Many organisations have complex internal administrative structures spread across the spatial dimension but for this case study (and presumably many other country-wide implementations of RHIS) the administrative structures follow the spatial dimensions as defined by our demarcation board and national government. Each province has a provincial administrative office for health, as does each district and sub-district. This geo-administrative alignment has simplified data management and warehousing by reducing the need to build complex internal structures (see Annexure D).

4.2.4 Patterns Emerge from a Prototype

An emerging theme relating to the horizontal dimension has been the need to relate organisational units to health service dimensions or classifications such as service type (e.g. clinic, community health centre, hospital, etc), administrative authority, as well as the rural/urban geographic setting. Because the management of dimension 'data' had been decentralized (underlying the inclusive and developmental approach adopted by system implementers) many of the design decisions were placed in the hands of information managers and information officers within each administrative region. Their data analysis needs were prioritised and back in the early 2000's the RHIS software tool DHIS v1.3 was evolving according to those needs. The following dimensions were prototyped and some were successful (later on became standardized in new iterations of the software) while others weren't (see Figure 19).

Location	No	Type	Category	Rural/Urban	Authority
Camdeboo		District Hospital	Public Health Facility	Urban	Provincially Aided Board
Camdeboo		Mobile Service	Public Health Facility	Rural	Cacadu District Municipality
Camdeboo		Satellite Clinic	Public Health Facility	Urban	Aberdeen Municipality
N Mandela Metro Sub		Mobile Service	Public Health Facility	Urban	Cacadu District Municipality

Figure 19. Different horizontal-dimensions were prototyped according to user-requirements in earlier versions of data warehousing.

These dimensions have stabilized over the past 15 years into super-classes. These dynamic dimensions were realised out of the prototyped successes from previous software iterations. What emerged from the prototype was the need for flexible dimensions and the only way to cater for this flexibility would be to design a classification system that catered for compulsory and exclusive settings. The 'orgunit group set' construct was created with the following options:

- Group Set Name;
- Description;
- Sort Order;
- *Compulsory* setting: a Boolean (yes or no) setting to determine if all organisational units are required to be allocated to a sub-group or not;
- *Exclusive* setting: a Boolean (yes or no) setting to determine if only one sub-group value may be allocated per organisational unit or no;

The 'orgunit group sets' that are permanent features across RHIS data sets are those that are compulsory and exclusive (see Table 8):

- *OrgUnit Type*: a mixture of hospital types and PHC service types;
- *OrgUnit Ownership* (authority): government (provincial or municipal), private or not-for-profit;
- OrgUnit Rural/Urban: Rural, urban or peri-urban;

PERMANENT HORIZONTAL DIMENSIONS					
TYPE	**Count**	**OWNERSHIP**	**Count**	**RURAL/URBAN**	**Count**
Clinic	3,202	Gov Province	4,902	Rural	4,078
Mobile Service	828	For-profit	835	Urban	2,715
EMS Station	486	Gov Municipality	562	Peri-Urban	42
Community Health Centre	267	Not-for-profit	272		
General Practitioner	258	Gov Other	184		
District Hospital	255	Province Aided	64		
Non-medical Site	210	Province EMS	13		
Pharmacy	191	Public Admin	3		
Correctional Centre	161				
Private Hospital	152				
Satellite Clinic	149				
Special Clinic	79				
Private Clinic	60				
Community Day Centre	59				
Regional Hospital	48				
Nurse Practitioner	41				
Reproductive Service	41				
Specialised TB Hospital	38				
Step Down Facility	33				
Oral Health Service	30				
Occupational Health Centre	27				
Specialised Psychiatric Hospital	27				
Health Post	24				
Mental Health Service	24				
Placeholder	19				
Provincial Tertiary Hospital	19				
Home Based Care	17				
Medical Centre	15				
Oral Health Centre	10				
National Central Hospital	8				
Specialised Hospital	7				
Forensic Pathology	7				
Environmental Health Service	6				
Hospice	5				
Frail Care	4				
Occupational Health Service	4				
Specialised Chronic Hospital	3				
Place of Safety	2				
Crisis Centre	2				

Industry Clinic	2
Rehabilitation Service	2
Health Education Service	2
Military Hospital	2
Midwife Obstetrics Unit	1
School Health Service	1
Health Sub-district Office	1
Ward	1
Specialised Orthopaedic Hospital	1
PHC Service	1
Specialist	1
VCT Clinic	1
Psychiatry Service	1

Table 8. Permanent 'Exclusive' and 'Compulsory' Horizontal Dimensions at Facility (OU5) Level.

The number and variety of level five and six organisational units has grown rapidly over the last few years. This has mainly been due to the need to understand health service utilization down to a (health) ward level especially in larger facilities (e.g. hospitals or larger PHC orgunits). Information officers at various levels are required to provide monthly and quarterly feedback reports containing aggregated routine data for entry in the DHIS. These feedback reports are specially designed per region and per health programme.

In addition there has been a range of new projects and initiatives aimed at addressing inequality and access to health services, all required to conform to this dimensional model. School health services are a recent introduction to the national health system as well as electronic medical record (EMR) data. The majority of routine data in the current system is collected through the tallying of paper register data (i.e. a paper register will record individual patient case information; the tallying or counting of aggregated totals is done as part of the data collection cycle by dedicated staff), while electronic systems allow for the automation of these data collection processes. The school health initiative requires the creation of an organisation structure that includes schools at a sub-district level (see Annexure C).

The massive environment that is modelled in the DHIS system is always undergoing changes. Reclassification of services, regional boundary changes, closing and opening of services and efforts to reduce health service inequality all lead to a dynamic set of organisational structure (see Table 9):

	EC	FS	GP	KZ	LP	MP	NC	NW	WC	TOTAL
1994	922	354	567	944	510	385	232	319	696	6923
1995	922	354	568	944	510	385	233	320	697	6928
1996	922	354	568	944	510	384	233	329	697	6937
1997	923	354	568	944	510	384	236	344	695	6955
1998	924	353	568	945	507	386	238	348	693	6960
1999	951	352	568	945	507	381	241	353	737	7034
2000	970	358	615	981	508	415	261	374	763	7245
2001	978	463	675	993	608	453	276	384	772	7603
2002	984	472	691	996	624	454	334	388	803	7748
2003	993	468	710	1014	641	456	334	394	809	7822
2004	1011	486	858	1031	648	463	336	412	830	8079
2005	1022	482	895	1047	660	468	334	421	868	8202
2006	1032	496	960	1070	662	472	336	427	874	8335
2007	1067	488	974	1087	674	475	341	454	921	8488
2008	1096	477	973	1110	683	477	343	477	930	8574
2009	1136	475	1006	1117	690	490	342	512	960	8737
2010	1152	468	1026	1113	701	501	343	520	965	8799
2011	1169	467	1053	1106	723	504	348	530	1032	8943
2012	1180	478	1046	1120	731	508	349	512	1037	8973
2013	1165	477	1051	1004	735	450	350	505	966	8716

Table 9. The number of facility-level organisational units broken down into provinces that were found in the DHIS between 1994 and 2013 (extracted from the DHIS system) with evidence of a clean-up between 2012 and 2013.

To summarise these developments in the health system model most of these structures had evolved in the following ways:

1. 1994 – 2000 (the early years):

   There was a need to develop a single version of the health system upon which routine data collection activities could be centralized. Together with the ‘health data elements’ model ‘flexibility’ became the attractor. Data customization and integration between parallel systems was possible that allowed for the presence of a single routine information system

2. 2000 – 2004 (the rapid growth phase):

The data model had grown to incorporate new sources of data (with a variety of vertical dimensions) and analysis perspectives were placing increasing pressure on the horizontal dimension classifications (e.g. administrative ownership of health services for accountability and rural/urban classifications). The data model had reached the limits of its hierarchical relationship abilities as the number of vertical levels was 'hard coded' to five.

3. 2004 – onwards (a new flexible model):

The new iteration of the data model was born on the successes and failures of its predecessor. Vertical levels were now unlimited but rarely went above six. Horizontal level dimensions were still stabilizing because of an uncertain health service classifications approach. Data model variations were tested but horizontal level dimensions would only become clear by 2008. In the absence of a health services classification guide the data model would adapt a 'group set' classifications approach based on the emergence of data analysis needs.

4.3 Evolution of Master Data Management

"There are approximately 41 patient-based systems, none of which talking to each other, but they are providing aggregated totals for output, that gets channelled into the DHIS. The DHIS creates the channel for integrating aggregated data from all these systems".

There has been much confusion around the classification of data within the RHIS environment. Different students of information systems that contribute to the development of DHIS struggle with the concept of information classification in which master data is treated as a special category of data. Based on the researchers experience within the developer community most prefer to simply go with the assertion that any data used to form links to operational or transactional data should be grouped under the umbrella term 'meta-data'. Very few go beyond this 'meta-data' classification but perhaps this stems from the perception that an RHIS is a data warehousing tool. This style of thinking is consistent with Kimball and Ross (2011).

Three major types of data have been identified in the RHIS system (DHIS) of South Africa:

- Master Data: Core & Relationship-based;

- Transactional Data: Routine;

In the case study core master data includes organisational units, data elements and indicators although a case could be made that indicators are relationship-based because all indicators are derived from data elements (data elements make up an indicator numerator or denominator formula). The relationship-based master data include organisational-unit hierarchy structures, data sets (collections of data elements), orgunit classifications (e.g. orgunit group sets and their associated orgunit groups), data element groups and indicator groups.

4.3.1 The Early Years

Right at the beginning of South Africa's new dispensation the health system was placed into a state of reform and the need for integrated and standardized health data was considered a top priority. A strategy was implemented to measure equity across health services and to pinpoint where resource efforts were most urgent (Braa, Hanseth, Heywood, Mohammed & Shaw, 2007). This change was measured by the creation and implementation of standardized systems for health data. At this time the development of a (data dictionary) tool to support and manage standardised data definitions was deliberated and the Australian (data dictionary) implementation was looked at. Eventually a different approach was adopted - one that followed an adaptive and incremental approach to data standards and information systems development.

In 1997 various initiatives were underway, all trying to address HIS fragmentation issues with different systems collecting different and (sometimes) overlapping data, resulting in parallel 'silos of information'. This problem is common in the absence of routine health information systems because each health programme within a ministry would be left to design their own information system instead of developing data elements as part of a unified system.

HISP was supported by the prototype software tool DHIS which allowed users to configure and collect data at various health facilities. Its data model also enabled data alignment and integration. Because of this flexibility health workers were able to configure their own data elements for health facility surveillance. An unanticipated problem with this flexibility was data fragmentation. Because so many variations of data elements were flooding the system the integration process became a data management problem, requiring the need for a single version of data element definitions. Negotiations began on the development of a minimal data set; this presented many challenges. Different health programmes could not agree on what should be excluded or included while different health facility authorities could not agree, as each had their own competing concerns in their respective facilities. Eventually consensus was reached, based on the reasoning that it was not possible to agree on everything but agreement should be reached on a basic minimum. The eventual minimum data set contained 47 data elements and is the first evidence of master data standardization in South Africa's new era of HIS reform. This initial incremental step enabled the first macro-level view of data across an entire province (Western Cape, soon to be followed by Eastern Cape) using a standardized set of master data.

During this time the DHIS software and its (master) data was under continuous development. An incremental leap forward for this initiative was the financial support and backing that came from a project known as EQUITY (a USAID funded initiative) that had an interest in data standardization in the Eastern Cape Province. The standardized 'minimum data set' was borrowed from the Western Cape and adapted to the Eastern Cape context and very soon it sparked the interests of other provinces who adapted the data set to their own contexts. All these incremental steps were enabled by a flexible software system designed to allow the customization and adaptation of standardized (master) data elements. It also was designed and supported by people who understood issues of complexity, data development and adaptation in changing environments.

Master data and its management was a growing concern due to the flexibility of the system and the complexity of the environment. With the establishment of a national 'essential' data set in June 2000, provinces were allowed to expand their data element collection with their own regionalized elements. The shared 'core' data set would remain untouched. Over the next few years many different data sets would be developed and piloted. The contents of the minimum 'essential' data set would be revised over scheduled workshops. Sometimes changes would be made to accommodate new health priorities such as the HIV/AIDS programme.

The National Health Information System for South Africa (NHISSA) committee was established in 1994 with the aim to enhance planning, management and evaluation of health services through the use of information ("Chapter 6: Health Information White Paper for the Transformation of the Health System in South Africa.", 2014) and in 2003 the idea for the creation of a national data dictionary re-emerged. A prototype was presented to NHISSA as an ICT tool to support the development and maintenance of the essential data set. It received no support because of a 'lack of collaboration and stakeholder engagement'. This is in line with the literature of Cleven & Wortmann (2010) on corporate MDM development and design.

Around 2004 the current version of the DHIS software (version 1.3), which had been designed to be flexible and dynamic, was placing restrictions on user requirements as more and more variations of data sets were being developed. It housed a master facility list that was regarded as stable by analysts from various other directorates within the health system. "We battled for a year to get consensus on a master facility list. We used the latest DHIS facility list at the time. That was the only recognized source of master facility data (even though it had flaws). Still it was the most consistent and well recognized source of facility data." The limitations of the data model were becoming evident. Limitations on the configuration and dimension-setting of master data made it clear that the prototype had achieved its objective and that it was time for the development of a more flexible version of the data model and software.

4.3.2 Expansion and Contraction of Master Data

By now MDM processes were evolving with each iteration of the essential data set (which took place on an 'adhoc' basis). Over the years certain patterns emerged relating to issues of master data management. In the absence of 'structures' or formalized processes for defining 'decision making data' different vertical levels were empowered to specify their own indicators and data elements for collection. Each provincial (health) office could make decisions on what data to collect without considering the impact or strain it placed on their regional systems (i.e. each level in the system inherits data-collection requirements from their parent-level).

Issues of variation and scale had to be controlled and coordinated. Data ownership was distributed, database changes were open to users across all levels of the system and master data challenges were being realised and addressed. Variations and growth of master data sets would remain a challenge. The number of data elements specified for collection had direct impacts on the work-load of facility level staff. Variations in data element 'naming' affected the integration process at higher levels. These complications were understood and slowly management processes evolved through needs-based implementation 'data workshops'. Over time various projects and programmes helped improve collective understanding of these challenges and processes were implemented to protect the integrity of data inside the DHIS.

These early phases of master data development gradually expanded until the routine data coming through the DHIS to regional levels was of such poor quality (in terms of completeness, timeliness and accuracy) that one sub-system reached a level of 'collapse'. In one extreme example of this 'collapse' an entire province of district, sub-district and facility level staff (involved in data collection and management) formed a province-wide coalition opposed to the extreme data-reporting requirements imposed by their provincial office. "When you do that without considering the facility level you're assuming the people at the bottom level have unlimited amount of time and they can collect an unlimited amount of data". They refused to continue and through a negotiation process resorted to only collect national-level data. Data quality had been so poorly affected that the provincial office had no choice but to agree to this extreme measure. This is one example of 'tightening' or shrinking of master data but what has been learned is that the master data development and specification process needs to be inclusive of all levels of the health system. Routine Health Information Systems are comprised of multiple levels working together to achieve health information goals. When one level acts on their own without considering the impact at lower levels – it can have terrible consequences for the integrity of the entire system.

4.3.3 A New Era for Health System MDM

2005 was a significant year for data management in the DHIS system of South Africa. It began with a critical change in focus for the essential data set. A shift in thinking that emphasised an initial focus on planning the output 'indicators' first resulted in the "essential" (input) data set being renamed to 'National Indicator Data Set' or NIDS. This came about by switching from thinking about what data 'could be collected' across health programmes to 'what could be measured'. This shift was a move towards planning the measurement outcomes rather than the data input definitions. This signified a major change in the use of information (Braa et al, 2007). No longer would the routine information system be burdened with data that supported the 'appearance' of productivity but was now enhanced with the foresight to plan master data with health-outcomes in mind.

4.3.4 Data Workshops: Self-Regulation Support Mechanisms

Over the years the DHIS had collected data representing thousands of health facilities across numerous data sets. Data fragmentation has been a continuous concern for managers and users of DHIS data. 'Data clean up and alignment workshops' have been critical for the regulation of master data integrity, relevance and reliability. Workshops were held on a regular basis - sometimes quarterly, sometimes on a needs basis. These workshops took place at provincial levels with the coming-together of district-level information managers. Very often provincial and district level information officers would end up with 'confused' data sets because of variations in master data names or definitions at lower levels. These clean up workshops would be supported by development partners (neutral parties) who would facilitate and negotiate the data standardization process.

Throughout these workshops data cleansing would be done for data element and indicator definitions with feedback being incorporated into the planning of future data set revisions. Data alignment and cleansing would also be done on organisational units (i.e. facilities) to incorporate new health service providers as well as manage any changes relating to classifications. Master data cleansing would be done out of necessity as fragmentation led to a decline in data integrity which negatively impacted the confidence of resource authorities and health programme managers. Over time these workshops would expose weaknesses in data flow processes leading to enhancements in the DHIS system and in the procedures associated with data flow. These workshops would also be used to support the development and training on new data specifications and definitions.

From 2006 onwards the problem of expanding data sets was beginning to stabilize. Self-regulation was taking place. After a year or two of expanding data sets provinces would 'contract' their data collection activities because of the work-load burden. Data sets would also stabilize due to policy or procedural changes. NHISSA began to provide constraints that limited this expansion. Health programmes wishing to increase data elements would need to get approval from the committee; this had positive and negative effects. Data quality would improve because of increased capacity at facility level but health programmes were overly cautious in their efforts to expand indicators, thereby reducing their understanding of health-problems. Learnings about what data could not be collected because of impracticality were also being shared. The impact of master data development practices was now being understood by development partners, funders and key DOH staff.

4.3.5 The DHMIS Policy: A Regulator of MDM and Health System Integrity

The DHMIS policy of 2011 (National Department of Health [NDOH], 2011) introduced a major focus on data quality assurance with focus on the development and review process of the 'national indicator set' and provincial level data sets. In order to ensure relevance of this crucial data the policy requires (master data) indicators and data elements to be reviewed as part of a scheduled planning cycle (known as the NIDS revision). This planning cycle is described as a structured consultation process that includes stakeholders from all levels of the health system, development partners and researchers. While the policy is already a few years old it has taken some time to formalise and structure the revision process. At present (2014) ICT tools are being developed to support both the information access component as well as a review tool. These include a national-level data dictionary to serve as a centralized reference point for master data definitions. This development has heralded a move from MDM level 1 (list provisioning) to level 2 (peer-based access) with elements of level 3 (centralized hub processing) (Tan et al, 2011) in preparation for the system wide move to a web based RHIS. This centralized data dictionary provides a variety of functionality to server the health 'enterprise' and provides the following master data components:

- organisational units across the various hierarchy structures (see Annexure C);
- data sets including associated data elements and validation rules;
- indicator definitions;

Other tools under development (in 2014) include a review tool as a subcomponent of the national data dictionary for initiating discussions around data definitions and includes a democratized approach towards master data development (data definitions can be voted upon) for programme managers and administrative-level managers.

4.4 Evolution of Data Quality

Data quality management as part of the BI "chain" in the DHIS has mostly been focused on the data input processes. Only much later on in its development would the links between MDM and DQ be realised.

Early on in the development of the DHIS (a data warehousing system designed to collect aggregate health system data) was it realised that data entry offered the greatest threat to data quality in the system. Data collection tasks were placed in the hands of nursing staff who were often heavily-burdened by large numbers of patients. They were expected to take on additional responsibilities such as those of data clerks. Data collection was done using various types of paper registers that would later (at the end of each month) be tallied and entered into the DHIS system.

Data values (counts of people) were tallied at each health facility on a monthly routine basis and were recorded against individual data elements or list items (see Figure 20) in the DHIS. The entry of these aggregated totals (numeric and sometimes large in scale) posed great risks for data quality accuracy. It was common for miscounts or keyboard errors to introduce unusual numbers. Also, it was possible for nurses to omit certain 'mandatory' values in data registers. These early learnings were realised and actioned through the development of data quality checks in the DHIS software.

4.4.1 Data Completeness

Pre-2004 data collection was focused on the entry of data arranged into specialised data files hosting a variety of data elements. Post-2004, with the advancement of DHIS software and its data model, these arrangements were grouped across data files and data sets. Each data set (a customizable collection of data elements) could be focused on a specific health area (or health programme) and comprised of anything between 25 and 600 data elements. Prior to 2004 much of the data quality completeness analyses were focused on the presence of routine data values only. Post 2004 the completeness tools expanded to include a variety of analyses that addressed measures for both data elements and data sets.

Data completeness 'compulsory' settings were created to assist users in the data entry process. Certain 'data elements' were identified as critical and compulsory for monthly reporting up to the national level. In the DHIS data entry saving of new values would be disallowed until all compulsory values were filled in or flagged with comments to explain their absence (see Figure 20). This completeness validation has remained an integral part of the system across all version of the DHIS software.

No	Data Element		Min	Max	Entry	Check	Comment
1	Nurse clinical work days	!	249	303	215		
2	Professional Nurse clinical work days	!	125	211	113		
3	Minuted meeting of committee / board during period		0	1	0		

Figure 20. Data entry screen with data quality ‘input’ measures: ‘compulsory’ setting (denoted by red-exclamation) to support completeness.

Pressure to ensure completeness of data resulted in the creation of tools to assist with the estimation and creation of missing values (an early effort to curb “incompleteness”). Some data elements held a higher ‘strategic’ value over others and their presence was regarded as crucial. Due to problems with uncontrolled expansion of master data (see Section 5.3.2) data quality completeness had to be addressed because of the need for a complete national picture of the health system. These tools became standardized inside the software and were to become known as “Missing Record and Outlier Analysis” functionality. This was addressed using statistical outlier analysis and completeness was seen as something that could be corrected by creating values based on historic trends. For over a decade data completeness remained a major area of concern in the system. Unknown to many stakeholders the office of the Auditor General (AG) would later (2008 onwards) go on to make a significant impact on data completeness by becoming a permanent participant and support mechanism in the evaluation of data for health.

4.4.2 Using Analytics Methods to improve Data Accuracy

From the start of the development of DHIS (1998 onwards) statistical methods have had a strong presence in the system (particularly in the DQ support tools). Because data was entered across time-ranges trends could easily be analysed using SQL queries to calculate averages, norms, acceptable ranges, etc. By using the standard deviation it was possible to determine acceptable value ranges (minimum and maximum ‘norms’) per health facility per data element. This was done by configuring two user-definable parameters for use during ‘save’ operation under data entry:

- *Period Count*: The number of data periods (months) used for the re-estimation of minimum and maximum range values (working backwards from the most recent month of data). Default was set to 6.
- *Standard deviation factor*: The number of standard deviations to be factored into the calculation for determining the 'maximum' above and 'minimum' below the AVERAGE value for the given period range per data element. Default was set to 2.

The calculation of 'acceptable' value ranges was easy to calculate and it soon became standard functionality to store profiles for facility and data element combinations. Nurses were assisted with these data accuracy checks. These acceptable ranges fluctuated as time progressed or as data values changed resulting in continually changing min/max ranges. The user base soon regarded these standardized data quality measures as 'min/max outliers' checks (see Figure 20). Data capture was continually presenting unusual values and soon 'outlier checks' were incorporated more permanently in the data entry process. As soon as entries were typed in after-update value checking was run. When an outlier was detected, a popup notification would appear requiring user verification to correct or re-estimate the acceptable value-range (see Figure 21). This example clearly makes use of analytics to enhance data input 'range accuracy'.

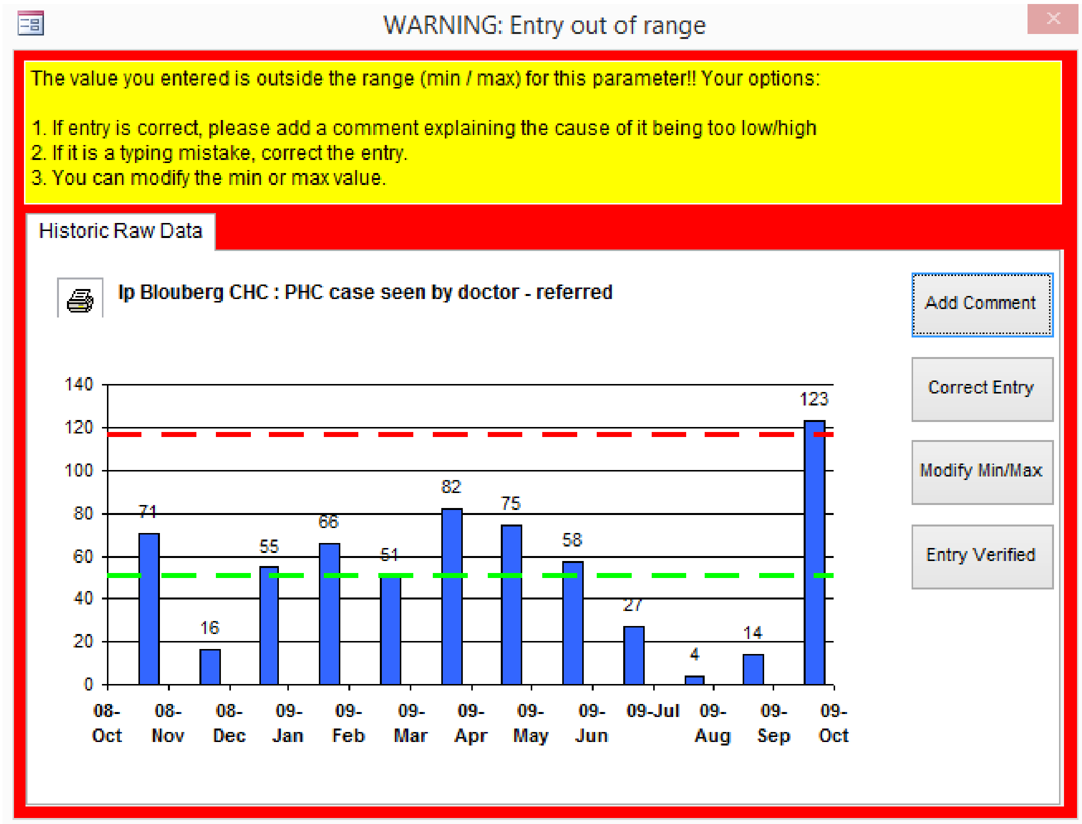


Figure 21. Example of a min/max outlier notification during data entry with the invalid entry represented by 09-Oct (acceptable value-ranges are determined using Standard Deviation).

Over time the data entry screen adapted to become more user-friendly. Once values are saved colour coding would immediately demonstrate min/max outlier values based on auto-calculated ranges (see Figure 19 for below minimum values highlighted in blue).

One of the limitations of calculating 'range-accuracy' is based on the fact that the range values are analysed longitudinally. A time-series analysis of the data used to calculate min and max values does not consider the possibility of seasonal trends (e.g. malaria occurs more frequently during summer months) or population migrations (e.g. visitors to coastal regions during summer months). Data modelling in the current system has not yet evolved to the point where analytics is sophisticated enough to detect or cater for such complexity.

4.4.3 Increasing Data Quality Depth with Analytics

Data entry quality checks were done using trend analysis which was based on historic data to determine acceptable value-ranges per data element per health facility. However, it was noted that relationships existed between different data elements, sometimes across data sets, which could be used for advanced analysis testing. This post -entry analysis was possible by taking advantage of logical 'relationships' that existed between data elements. Data elements represented different cohorts or population groups and this provided opportunities for the creation of logical tests between similar sets of values. Standardized tests were developed which became known as absolute and expert validations and were based on a simple form of pattern recognition (see Table 10). Because most data (values) were assumed to be stable (with the exception of epidemics or outbreaks) fluctuations would only occur under exceptional circumstances or in the presence of poor data.

The creation and execution of these data-validations allowed users to either:

I. identify outliers or abnormal patterns in order to make corrections;
II. identify outliers or abnormal patterns and allow the user to (try to) explain why the abnormality occurred by entering a descriptive comment.

Absolute validation rules applied to situations where one value could not be higher than another. An example would be the sum of child-attendance vs total attendance. The child headcount (sum of children under the age of five years visiting a clinic) could not be higher than the total attendance (total headcount of adults and children) at the same facility (unless an error was introduced). These 'obvious' validation rules were more easily defined but a more complex set of validation rules existed.

Expert validation rules (also referred to as statistical rules) were designed to be more flexible and test ratios between data elements. These validation rules tested correlations between data elements, e.g. data element 'children with diarrhoea' should correlate with 'child attendance'. If the total headcount for children (visits) goes up, one would expect a proportionate or similar increase in the number of child diarrhoea cases. These expert rules were dependent on the knowledge that came from years of experience as a nurse or health-professional. Only those 'experts' would understand the correlation between different sets of data-elements and were named expert-validation rules.

These validation tests followed a pattern check and were used to identify anomalous outliers. Guidelines for the development, application and use of these different validation rules were made standard-practice (see Table 10 guidelines) and incorporated into the training manual.

Good quality data management entails monthly review of data before it is submitted for data entry. Scrutiny of the data should identify grossly incorrect values. In addition, facility managers, and supervisors should: a) determine 'Absolute Validation Rules' to ensure that poor quality data is identified and corrected immediately after data entry; b) determine 'Expert Validation Rules' to help identify deviation from normal trends. Expert Validation Rules therefore act as pointers of possible discrepancies in the data or increasing incidences of diseases.

Table 10. Extract from the DHIS training manual of 2003 referring to guidelines on the use of validation rules.

Around 2004 different projects began highlighting issues of data quality coming from paper registers, one example being a "hospitals" project. By this stage data collected through the DHIS system was used by provincial and the national departments to manage the allocation and reallocation of resources across the entire health system. In one extreme case of poor data quality, a hospital was identified by the national department for possible closure. According to data collected through the national DHIS "stream" it had a consistently low 'bed-utilization rate' (see Annexure B) together with a high 'cost per patient-day rate'. Accepting this data as accurate meant an interpretation of wasteful and unnecessary expenditure, prompting a decision to close the hospital and reallocate resources elsewhere. When this decision was made known it sparked an outcry and an intervention study was conducted (see Figure 22). This resulted in the first known data quality audit that disproved a health management finding due to poor quality data. The hospital had reported an extremely low 'bed-utilization rate' because it had poor DQ completeness. Nearly half of all data was not collected or captured in the DHIS.

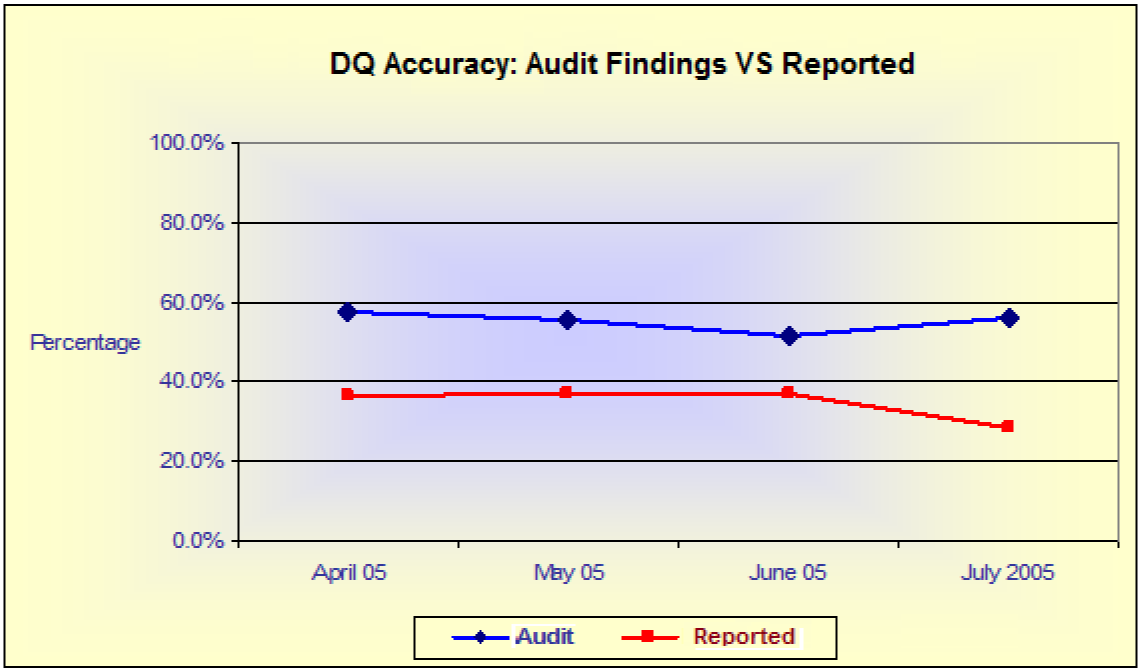


Figure 22. DQ Audit findings comparing what was reported VS actual data.

In a separate incident one hospital manager admitted to purposefully inflating data values because higher head-counts equated to higher budget allocations.

4.4.4 Data Workshops: Regulating Mechanisms of Data Quality

For many years data clean up workshops were used as mechanisms to regulate DQ in the DHIS system. As already stated (see section 5.3.4) these workshops were attended by various participants such as District Information Officers (DIO), Provincial Information Officers (PIO) and facilitators from development partner organisations. In terms of transactional (routine) data quality accuracy and completeness evaluations have been addressed in these workshops using tools developed inside the DHIS software. Examples include “Missing Record and Outlier Analysis”, “Absolute and Statistical validation rule checks”, Data Completeness reports (see Annexure E) and occasional eye-balling (pattern enquiry) of raw data. These various DQ tools (based on analytic methods) have supported data “clean up” through inclusive workshop situations.

4.4.5 DQ Enhancing Initiatives

Between 2004 and 2008 numerous data quality problems were understood and these learnings had heavy bearing on the establishment of the "National Data Managers" project of 2009 which had the following DQ objectives (sourced from unpublished internal project documentation):

- Ensure efficient data flow to the NDoH *(timeliness)*
- Provide regular reports and feedback on data quality and completeness
- Improve data completeness and quality *(improve overall quality)*

This project provided a deeper focus on data management and quality which strengthened awareness of DQ across provinces. The learnings from this project influenced the development of yet further support-tools in software and DQ assessment processes. Examples of software tools include:

- *Snapshot summary report*: this report was run for any region or level of the hierarchy (for a data set). Data values were assessed for *completeness* as measured by 'expected' (count of values), 'actual' (how many are present) and 'missing' (the difference). These variables allowed for the calculation of a 'reporting rate' (percentage). These snapshots would be taken for a time-range and could be compared retrospectively to check how data values changed over time (see Annexure E screenshot of tool). It also included a data-variability of values sections.
- *Data completeness report*: this report was run for any region or level of the hierarchy and could be used to output a variety of metrics related to completeness (1's and 0's for missing, number of records captured, % of data set captured, % of orgunit profile values captured, etc). These outputs were listed per orgunit per month.
- *Data timeliness report*: this report was run to produce metrics related to the data-capture process. Because data for a month is generally collated after the month-end, timeliness could be measured as the number of days after the end of a month for the following metrics:
    - *Initial Capture*: the measure of days between end-of-month and earliest (date-stamp) value captured for a data set
    - *Last Edit*: the measure of days between the end-of-month (date) and the maximum edit-date for all records in a data set
    - *Edit Period*: the measure of days between [initial capture date] and the [Last Edit] date for which values changed in a data set

As a data quality measure *accuracy* has consistently eluded system administrators because developing of a self-auditing culture has not happened naturally. Independent audits have been the primary method for obtaining DQ accuracy measures but this has mostly been done on random facilities. Between 2009 and 2011 different data quality audit (DQA) tools were developed with low user uptake and since 2009 the office of the Auditor General (AG), employed through Treasury, has provided audit results that are used annually for the assessment of data quality. The AG has therefore been instrumental in highlighting issues of data quality through their participation in the auditing of health information systems. The outcomes of their audits may be “qualified” or “unqualified” in which case they are supportive of performance plans or may result in “sanctions” being imposed for poor data quality providers (such as slashed budget allocations, which are negative feedback loops as under resourced health services often lead to further problems in management and delivery of health services).

From 2010 onwards a more stable and complete set of data was being used for the planning of budget-expenditures (provincial administrations could use different indicators to determine expected work-loads across primary health care services for different regions). Now that the AG was a participant in the assessment of data quality, they proceeded to conduct audits of data collection systems providing outputs to Treasury. One of their major concerns was around the security of the master data elements inside the DHIS. The open-source/open-content philosophy of empowering data users at all levels of the system (national, provincial, district, sub-district and facility) had become a fragmentation risk affecting integrity of the overall system. When Treasury of South Africa (the financial resource allocation authority) ran the risk of misallocating funds that run into billions of Rand’s the need for quality data became an urgent priority for the NDOH. Not only was data being submitted late but master data fragmentation was slowing down access to management information.

It was too easy for system users to modify master data and create data management and integration problems. The AG highlighted these data management concerns along with the need to improve timeliness. Based on their findings changes were implemented to cater for the locking-down of specific data elements which centred on the core national data set (see Figure 23). A balance had to be reached that satisfied security concerns but one that also maintained a level of openness for other master data sets. The solution was to password protect each database file with an encryption ‘key’ that was known to administrators and database managers. In addition to securing back-end databases each master data record required an optional password-protection 'key' to prevent uncontrolled changes in the front-end. In addition to the national-level 'lock-down' of data sets an 'editing window' configuration setting was added.

When presenting budgets to Treasury many provincial-administrations would find themselves using data that was different to the values arriving at the NDOH. This problem resulted in the adaptation of data element design. The 'editing window' setting was implemented to prevent changes to values already submitted through the DHIS system older than the specified time range, e.g. an editing-window value of +3 would allow the editing of data values captured in January up until the end of April. It would be locked from 1 May onwards. These changes were implemented in 2011 and saw an improvement in data quality for provincial and national administrators. Not only was integrity of data being protected but it also reduced data management problems for national-level data sets. Turnaround time (timeliness) was also being reduced.

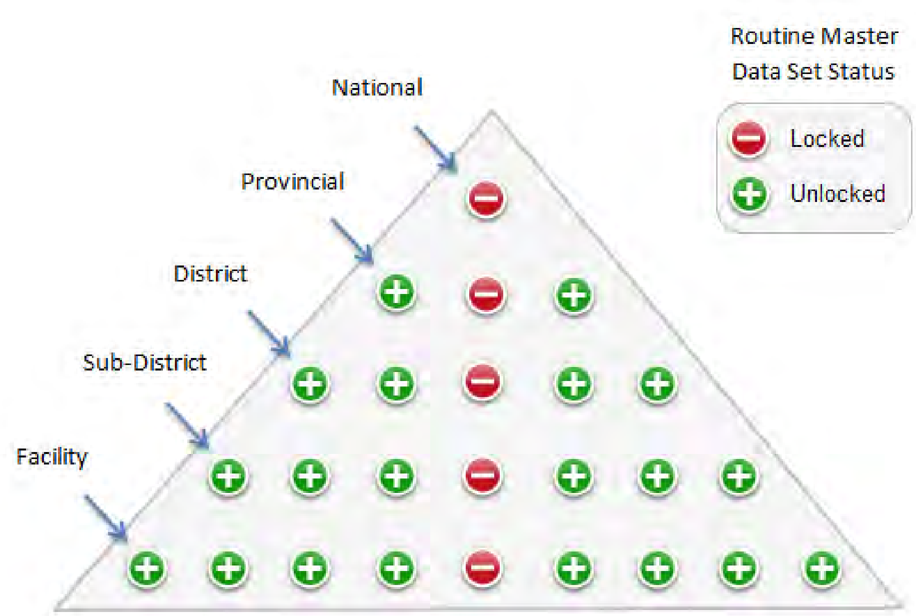


Figure 23. Different vertical-levels specify their own data-sets to be collected by the DHIS (the core national-level data set is 'locked' to protect master-data integrity).

Other strategies for improving data quality 'accuracy' have been implemented. The submission of routine data (i.e. collection activities related to the tallying of registers, capturing of totals in DHIS and submission of export data to regional levels) would normally take place once off at the end of the month. Efforts to move routine data collection to a daily schedule are currently underway (in 2014) with high hopes that it will lead to an improved data quality across the system.

Other efforts include initiatives aimed at reducing the workloads of nurses and data entry clerks. At present a huge number of paper registers are expected to be filled in each day in PHC facilities. This places a massive burden on staff that are expected to fill in duplicate and sometimes triplicate copies of data. Projects are underway to pilot new 'rationalized' registers with the hopes of relieving this workload, which may lead to improvements in data quality.

4.4.6 The DHMIS Policy: A Regulator of Data Quality

The DHMIS policy of 2011 (National Department of Health, 2011) has introduced a significant focus on the DQ component of BI. The policy makes direct reference to the South African Statistical Quality Assurance Framework (SASQAF) and refers to eight dimensions in the following order: relevance, integrity, timeliness, accessibility, reliability, completeness, accuracy and coherence & comparability. The policy describes these DQ dimensions with implementation guidelines:

Relevance is addressed as part of the MDM (NIDS and PIDS review) cycle and relates mostly to the development and specification of health-indicators and associated data that is collected through health facilities. The policy stipulates the relevance of master data being assessed in a two year cycle. 'Infrastructural' master data (organisational units and their associated hierarchies and classifications) are reviewed on a continuous basis.

Integrity is addressed by maintaining values and practices to ensure confidence in the system and in the production of health information. This infers adherence to a) defined integration processes and cycles (see MDM section 5.3.5), b) the screening and exploration of data quality and integrity checks built into the RHIS, and c) locking of data beyond the editing-window period as defined in SOPs.

Timeliness is addressed through stipulations requiring data to be submitted according to timelines set by the NDOH. A timeliness analysis report is also required to be developed and submitted on a quarterly basis.

Accessibility is addressed through stipulations requiring information users (which include health programme managers, national-level management, provincial-level management and district-level management) to have access to information as well as a medium through which information can be accessed. This component includes a stipulation making the assessment of DQ completeness and overall quality a responsibility of the users listed above.

Reliability is addressed through stipulations requiring provincial level information units to identify and review organisational units that provide low quality data and assess the implications on poor health service delivery. This stipulation also requires a follow up investigation at each organisational unit to determine the factors leading to low quality data.

Completeness is partially addressed through the stipulation that national, provincial and district level processes shall be implemented to test and verify data consistency and completeness. This section also stipulates that district health managers, hospital CEOs and health facility managers are to have data quality timeliness and completeness outcomes drafted into their performance contracts.

Accuracy is addressed through stipulations that require health facility managers to conduct accuracy assessments before submitting data for capture. These assessments include the application of relevant data validation rules (see section 5.4.3) which are made available inside the national data dictionary. In addition facilities are required to conduct their own data quality audits with findings reports and improvement plans.

Coherence and comparability refer to the ability to bring together data from different sources that represent the same characteristics for places at the same point in time. The RHIS of South Africa (DHIS) should be compared with survey data from time to time.

All these developments have changed the DQ landscape of the case study. From 2009 to 2010 the information quality level 1 (ad-hoc) took shape through the national data managers project (see section 5.4.5) which culminated in significant learnings that were incorporated into the DHMIS policy of 2011. While the specifics of level 2 are still being realised, e.g. information products (IP) are still developing, what is worth noting is the information quality (IQ) requirements have been specified. This policy has put in place requirements that when fully realised should move the DOH from level 2 (define IP and IQ) to level 3 (IQM initiative) or level 4 (IQ assessment).

4.5 Evolution of Data Warehousing & Analytics

The DHIS system has evolved its analytics capabilities since around 1998 when its first implementation took place in the Western Cape. At the time data collection resulted in large amounts of aggregated values that were analysed using complex queries. Over a short while these queries evolved into standardized aggregation rules comprised of numerator and denominator formulae. Over two iterations of software the data constructs would remain largely the same. Data elements would represent counts of health phenomena (e.g. headcount of visits to a clinic) and indicators would represent a calculation based on inputted data elements, e.g. utilization rate would be the sum of headcount visits divided by the local population estimates. Data collection would take place within a data file (see Figure 24) with ETL processing creating or updating a paired data mart file. Most of the improvements between DHIS version 1.3 and version 1.4 were to address relational database design issues. The version 1.3 made use of a relational database design but its primary and foreign key fields were text based which were inefficient and outdated. This upgrade applied to both the data file and data mart designs with other enhancements addressing limitations to the data model.

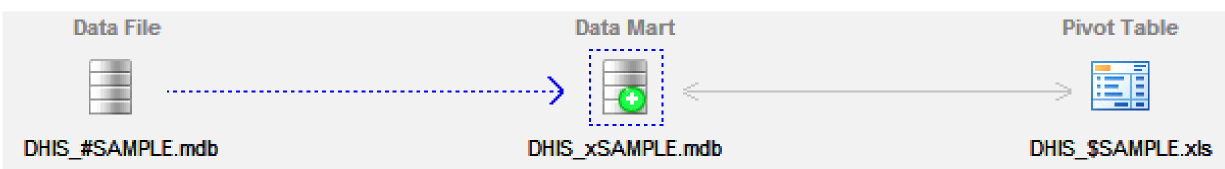


Figure 24. Data is transformed using ETL processing and migrated to a data mart from which pivot tables are refreshed.

Except for new analytic requirements in the NHIRD environment the ETL processing in the DHIS remained mostly unchanged since the early 2000's. It was adapted slightly to cater for new data dimensions and was even adapted to support a parallel process that generated data quality metrics in data quality marts. The Indicator 'construct' has persistently formed the basic measurement 'tool' of the DHIS system (see Figure 25). It is comprised of a numerator and a denominator, each able to be represented in a math formula based on data elements collected in data sets (see Annexure A). It represented a ratio, percentage, rate or proportion of one variable over another (numerator divided by denominator) where either one can be comprised of a complex combination of data elements or numeric variables.

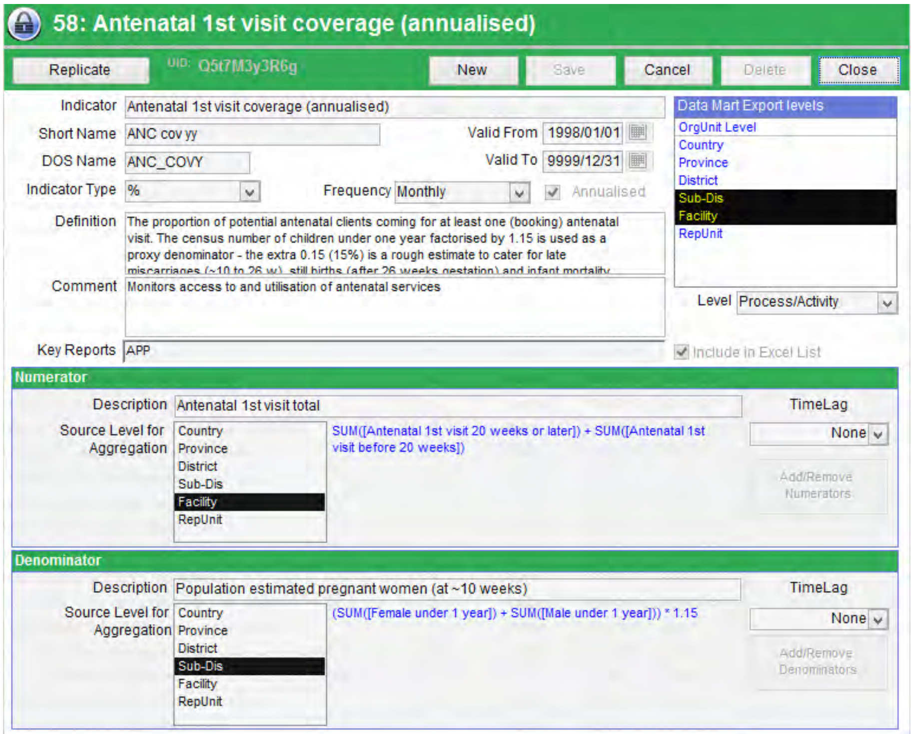


Figure 25. Indicator specification screen from DHIS v1.4.

4.5.1 Distribution of Data in Spread Marts

The data mart design is relational and supports OLAP querying. During the 'export to data mart' (ETL) process all values collected at facility or organisation unit level 5 (OU5) get aggregated up to sub-district level (OU4), and the process is repeated until the national level (OU1) is reached. The same process applies for indicator calculations where numerator and denominator values get aggregated up. The resulting data mart file is linked to MS Excel through multiple pivot tables which are refreshed via ODBC links.

Since the early 2000's DHIS data has been collected, integrated and redistributed back across all provinces through a distribution process that includes data files, data marts and Excel spread sheets. (see Figure 24). These 'spread-marts' have been extremely popular with users over the years as they are portable and comprehensive in content. Statistical analysis has almost exclusively been accomplished using pivot tables and the processing tools with MS Excel. This has been supported by regular training workshops on 'use of information' as well as the annual 'UWC Winter School'. The 'UWC Winter School', hosted by the Faculty of Public Health, has provided training since the early 2000's on various DHIS courses ranging from introductory, to intermediate and advanced courses. Much of the coursework has focused on the use of pivot tables for developing reports. This annual event has helped develop the analytics capabilities of hundreds of different participants over the years, many of which were from the DOH.

The use of DHIS data within the DOH has mostly been through the application of pivot table functionality with feedback being provided to health programme managers and regional administrators in order to comply with statutory reporting requirements, e.g. the Quarterly Reporting System (QRS), implemented by Treasury for monitoring and evaluation purposes, and for Annual Perform Planning (APP). Analyses have also been done to support local knowledge of populations in order to match health services to supply and demand. "Population is really important, where must health services be provided and also allocate resources to serve that population"

4.5.2 The Move to Self-Service BI

South Africa is a developing country and as such its ICT infrastructure is under developed. Due to its limited internet connectivity web-based information retrieval had not been prioritised by the DOH. However, the development of a centralized web-reporting system began in 2007 as part of a small contract with the GPDOH. Up until then all nine provinces were comfortable using spread mart data (distributed on a monthly schedule). GPDOH were first to proceed with an investment that supported their existing GIS capabilities. Because they already had a significant momentum with their in-house GIS solution they recognized the need to make data available in an online format. This project coincided with the upgrade from DHIS 1.3 to DHIS 1.4 which saw various improvements to the data mart model. The specifications for their web-reporting system were finalised and the reporting-solution was developed over a six month period. After a cooling-off period a second follow-up project was funded to conclude partially completed functionality and to enhance existing output formats. The resulting capabilities included:

- Scrolling marquees depicting indicator upward or downward trends (see Figure 26);
- a dynamic charting that covered both indicator and routine data across multiple vertical levels;
- dynamic reports across horizontal and vertical levels with OLAP (drill-through);
- dynamic population pyramids (population data is a basic requirement in any RHIS);
- user definable thresholds for different indicators (utilizing colour coding for 'outliers');

- Dashboards that integrate a variety of prebuild output formats on a single page;

Most analytic capabilities of data users were being utilized in the spread mart environment. Not much learning was adapted to the web environment except for a basic pivot table 'like' outputs. Beyond that the self-service system was completely fresh to users and it was very well received by management and IT (see Annexure F). Different interactive variations of outputs were developed over the next few years with the majority of tools being developed as flexible and generic rather than as standardized reporting templates. The population pyramid was adapted to include ratios of population to organisational unit counts (see Annexure G) while the web tool had to be redeveloped to cater for new types of web browsers (e.g. iPad, mobile browsers, etc).

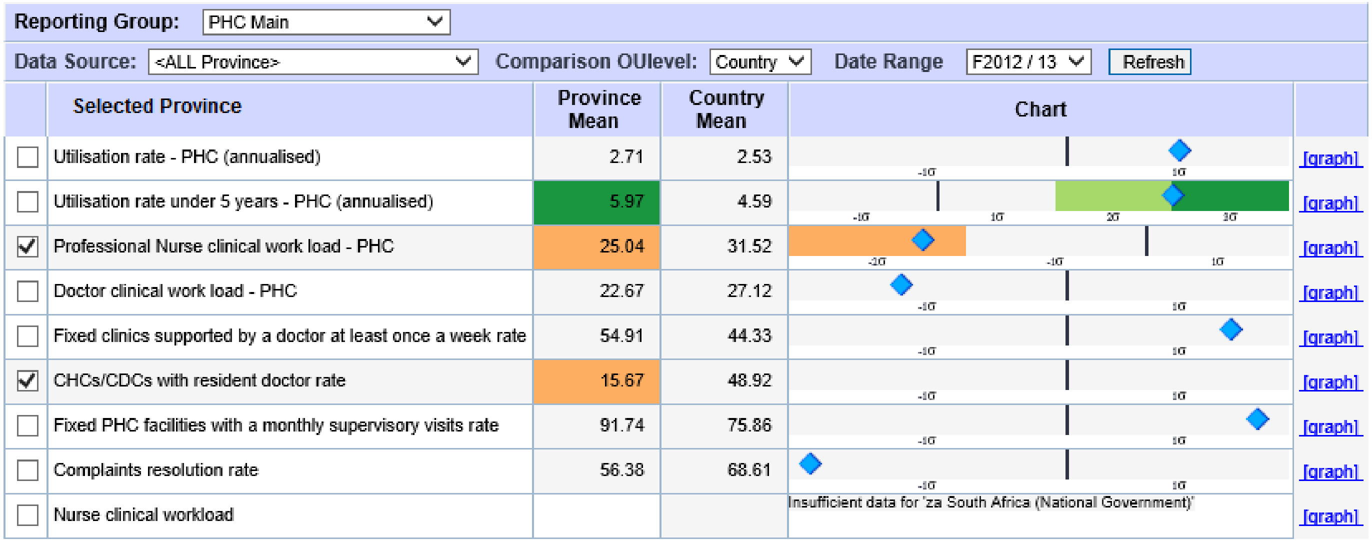
4.5.3 Analytics Leverage Point: Indicator Slope or Polarity

In the case study the conceptualisation of an indicator "slope" or "polarity" setting could be regarded as a critical leverage-point or stepping stone in the use and presentation of health data. The ability to define this attribute for an indicator has led to the development of new information presentation outputs, and new applications for analytics. Discussions around the concept began in 2008 at a conference in India but arriving at an agreeable definition for the concept has proved to be a challenge. In business or economics growth between two points, i.e. slope, is most often a desired outcome because of the perceived increase in value (more money is a good thing). In "health" we find ourselves working with complex variables. Across health programmes a wide variety of outcomes or measures are recorded for analysis as indicators. The upward trend of a slope (i.e. growth) for an indicator is not necessarily a desirable trend for health managers. For some indicators an upward trend is a negative outcome (e.g. infection rate). For others it is desirable (e.g. follow up visit rate for infections) while for some slope direction has no real perceivable meaning (e.g. wheelchair issuing rate has little meaning to the majority of information users). It could be argued that neutral (polarity) indicators have no real use in an HMIS so this could assist with the clean-up/reduction of master data definitions in a RHIS.

Figure 26. A scrolling-marquee with indicator values and growth-trends.

The application of this indicator setting has had a tremendous impact on a user’s first experience of the web-reporting system. It has enabled the color-coding of output data and has led to the development of next-level analytic tools. Examples of these outputs include a standardized ‘marquee’ toolbar (see Figure 26) a performance-gauging report for indicators (see Figure 27) and experimental graphic outputs (see Figure 28). These outputs are used to assess performance of indicators between different periods.

When the NDOH established the need for a centralized warehouse for health data, they were presented with the GPDOH solution. They immediately requested support from HISP to begin the national-level implementation. During the early phases of this secondment, a visiting doctor from a UK public health observatory presented their analytic work to the NHIRD team. Their work utilized statistical methods to plot over or under performance using standard deviations. Other tools included a Z-scores hypothesis testing tool that could be used to test correlations between pairs of indicators, (e.g. was there a significant relation between areas with poor sanitation services and health facilities with high diarrhoea incidence). The presentations of these two tools led to new outputs inside the NHIRD reporting system becoming standardized tools for assessing performance, affectionately referred to as the ‘performance manager’ (see Figure 27), while the Z-score tool would remain a desktop utility and be used in the hypothesis testing of a range of gender specific health assessments. None of these standardized tools would be possible without a ‘polarity’ variable to represent a desired outcome for each indicator.



Reporting Group: PHC Main

Data Source: <ALL Province> Comparison OUlevel: Country Date Range F2012 / 13 Refresh

	Selected Province	Province Mean	Country Mean	Chart	
☐	Utilisation rate - PHC (annualised)	2.71	2.53		[graph]
☐	Utilisation rate under 5 years - PHC (annualised)	5.97	4.59		[graph]
☑	Professional Nurse clinical work load - PHC	25.04	31.52		[graph]
☐	Doctor clinical work load - PHC	22.67	27.12		[graph]
☐	Fixed clinics supported by a doctor at least once a week rate	54.91	44.33		[graph]
☑	CHCs/CDCs with resident doctor rate	15.67	48.92		[graph]
☐	Fixed PHC facilities with a monthly supervisory visits rate	91.74	75.86		[graph]
☐	Complaints resolution rate	56.38	68.61		[graph]
☐	Nurse clinical workload			Insufficient data for 'za South Africa (National Government)'	[graph]

Figure 27. The 'spine chart' or benchmarking tool displays regional performance against the parent-level averages supported by the use of standard deviation and indicator 'polarity' for colour-coding.

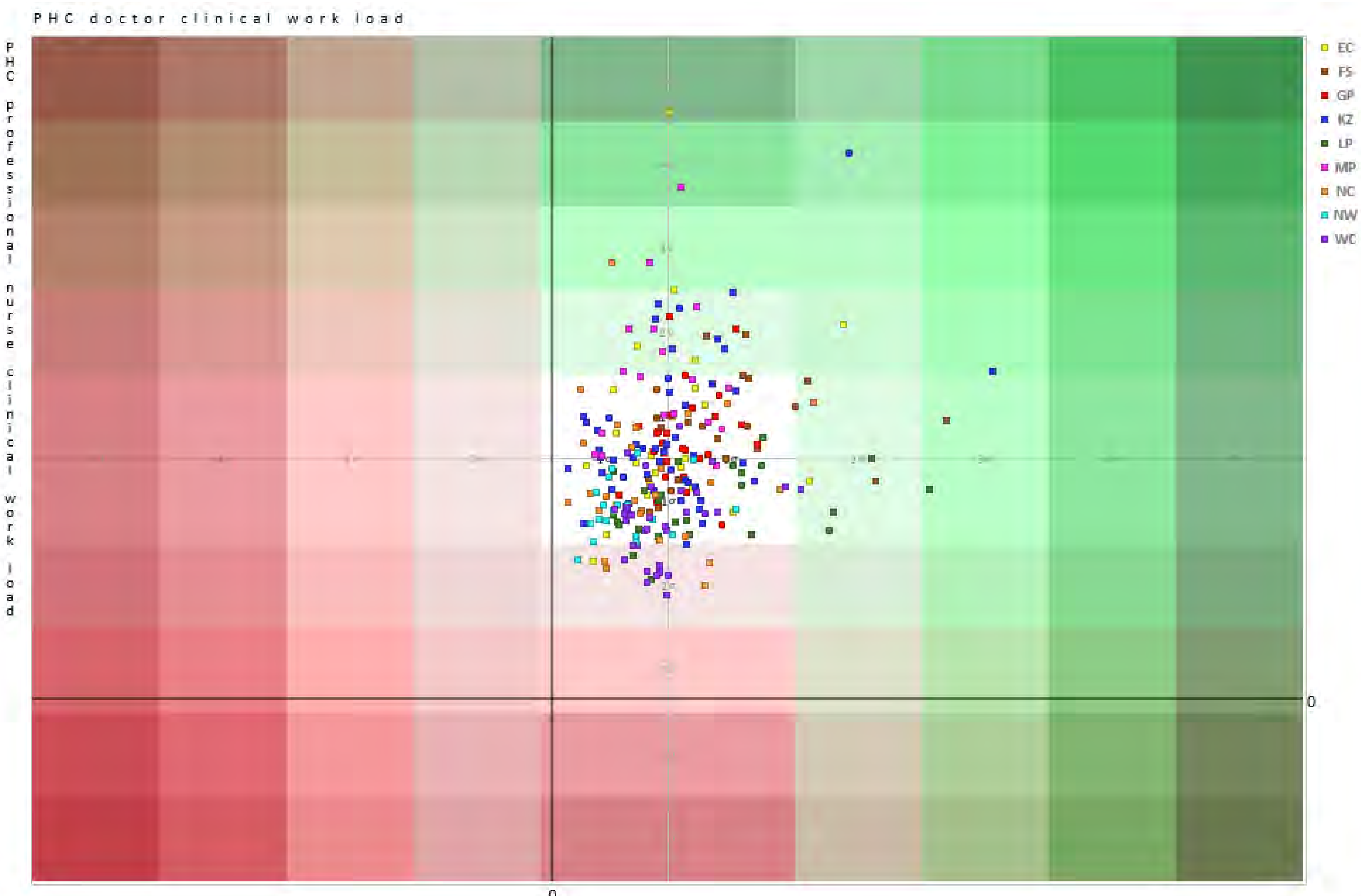


Figure 28. Experimental scatter-plot chart where two indicator variables are compared at sub-district level represented as colour-blocks (red-zone denote under-performance with each block representing a single measure of Standard deviation).

5. Discussion

The following is a discussion on the two major subject areas of this research project, namely the environment (as a complex adaptive system) and the emergence of self-service BI in public health. This will be followed by a summary of key learnings. The discussion will start with a review of the biggest challenge faced by the researcher: the development of ontology to support the research process.

5.1 The Ontology Development Challenge: BI and the area of HIS

The initial concept for the research project was an investigation into the establishment of the NHIRD as a BI platform for 'health'. However, during the development of a BI ontology to describe the platform critical underlying processes emerged from the case study supporting the literature of Tan et al. (2011) on enterprise level business intelligence. This prompted a deeper review into literature on MDM, DQ, Analytics and Data Warehousing as sub-areas of BI. A similar process was followed with regards to the HIS domain. An HIS ontology was required in order to interpret the underlying systems that provided data into the NHIRD platform. Literature from Lippeveld (2001) provided guidance for the domain classification of the DHIS as an RHIS system within the domain of HMIS which is located within HIS.

5.2 Case Study Environment as a CAS

Complex adaptive systems are guided by unwritten rules of interaction and feedback loops among agents and groups of agents that learn from their interactions. They are constantly changing and adapting their internal rules within their environment. CAS theory has enabled the researcher to observe some of these 'patterns' and perceive the RHIS in the case study in an evolving state at different times and from different perspectives (i.e. the different BI focus areas). It cannot be described as a fixed or static representation of processes and people but rather as a system under continuous change as a result of iterative learnings.

6.2.1 First Steps out of a Negative Plan and Control Structure

CAS theory refers to incremental learning points as the evolutionary steps that move a system forward while leverage points appear to be those points of control or power that govern processes or rules of interaction. Sophisticated hierarchical structures appear to be governed by top-down rules of interaction while a CAS has a moderate or flat hierarchy which provides many opportunities for interaction and learning. In the case study environment there was a dynamic interplay between agents in a zone of complexity (see Figure 29) which was encouraged through sustained efforts by development partners who understood this need. This helped encourage interaction, transformative feedback loops, distributed control, growth and evolution that would not take place within the zone of "plan and control".

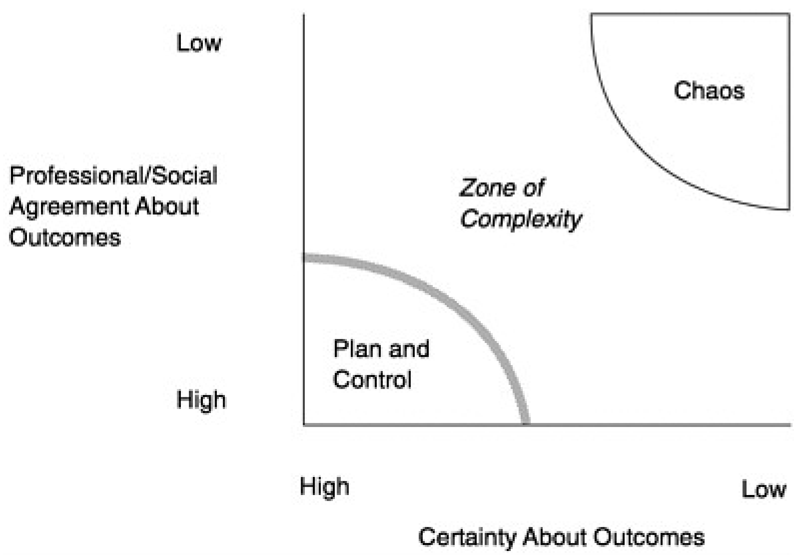


Figure 29. Zone of complexity (Stacey, 1996).

The environment in which the routine health information system (DHIS) emerged as a tool to support HMIS development comprised of hundreds of people representing various types of organizations with different professional backgrounds. The DHIS software, its master data content, its data flows, data quality procedures and data outputs evolved over time beginning with the establishment of relationships between agents within the same public health system. One of the biggest impacts of Apartheid was the restriction of information flows and the development of relationships between people of different backgrounds, e.g. the City of Cape Town's health administration was segregated from the Cape Metropolitan's health administration because of an accumulation of rules or policies. The stringent top-down control mechanisms of Apartheid negatively impacted naturally occurring information flows and interactions.

From 1996 to 1998 foreign development partners would experiment and prototype different data collection tools until the DHIS version 1.3 was established and adopted in the Western Cape. This prototype allowed routine data to be integrated into a data warehousing environment that provided regional overviews for health workers. This realisation had major effects on the local system. It helped initiate dialog between different health administrations that were historically alienated from one another. The realisation of the power of information for decision making acted as an attractor that led to the creation of relationships and new information flows. People from different sub-districts and backgrounds began participating in the development and discussion of health data. This unifying process helped develop relationships and trust among system agents. By 1998 the evolving data warehouse was under rapid development with support from international funders (USAID), researchers and academics from different Universities (Oslo, UWC and UCT) as well as different health administrations who all participated in the evolving content and structures. These changes reinforce CAS characteristics identified by Hammer et al. (2012): from continuous varying interactions characteristics 'Local & Remote', 'Non-Linear Interactions', 'Continuous Interactions', 'Connected Open Systems ' and 'Relationships co-evolve'; from patterns development characteristic 'Patterns and Attractors' ; from people factors characteristics 'Histories' and 'Space possibilities'.

By 2000 continuous varying interactions were taking place regularly between large numbers of people. Managers and administrators were developing focused data sets for reporting to health programmes and regional administrations. Their expanding data sets would test the limits of system capacity as well as boundaries. New reporting requirements would propagate outwards until secondary reactions from a range of agents, e.g. data clerks, data managers, programme managers, academics and public health specialists, would result in feedback being transmitted back to originators (e.g. the need for a rationalized master set of data definitions). These changes reinforce CAS characteristics identified by Hammer et al. (2012): from continuous varying interactions characteristics 'Positive & Negative Feedbacks', 'Large Numbers' and 'Rich Interactions'; from patterns development characteristics 'Patterns Emerge' and 'Origins of patterns'; from people factors characteristic 'Space possibilities'.

5.2.2 Whole System Ignorance through Non-Linear Interactions Create a Negative Feedback Loop

In an example of whole system-ignorance with a negative feedback loop provincial authorities made a series of decisions that had negative impacts for the local system which propagated outwards affecting the whole system. The unpredictable decision to approve all data collection requests by health programmes, researchers and administrators without following a collaborative data development approach (i.e. shared control and decision making power) resulted in the rapid expansion of master data which would ultimately result in a negative feedback loop. Unaware of the impact of their decisions managers created “unrestrained lists of health data definitions” to be collected province wide by facility level staff. This data was also required to be managed by sub-district and district level information officers. The scale of their actions was not understood at the time. The rapid expansion of data elements had tumultuous effects. It resulted in an overloaded system breaking down within two years. The impact was a system wide imbalance which led to a breakdown in local relationships and a drop in overall data quality. Data collection tasks were heavy and out of balance resulting in incomplete and poor quality data which had an impact on the national level. Because data coming through a province was of poor quality the national totals and averages were negatively impacted. The end result was a major drive towards an ultimate but simplified data set. The province dropped all data reporting requirements except for those required by national (i.e. the NIDS data set). The rapid expansion of data tasks caused a partial collapse of the local system and a minimalist drive from chaos to control would be necessary to stabilize data flow processes and improve overall data quality for the entire system. This learning point meant that future reporting requirement would be adjusted to consider load balancing for the system.

5.2.3 Self-Organization Develops from Internal and External Factors

Because data was being collected each month the exchange and integration process was repeated twelve times each year (per district office, per provincial office and for the national level). This offered numerous opportunities for learning. Data was being collected within the DOH by DOH staff but with support from a wide range of agents. Different agents with different backgrounds all felt accountable for the processes, software tools, databases and outcomes of data collection and integration activities. The system was open and interactions would take place crossing regular ‘employee’ boundaries. It was common for DOH employees to engage with non-DOH employees. Information exchange took place in the form of data management workshops, data clean up workshops, winter school training sessions but most frequently over the phone. These routine interactions and exchanges allowed for dynamic and rich information exchanges and learnings that were adapted or shared with other CAS participants. Examples include software developers engaging with information officers, health programme managers engaging with data managers from NGOs and CEOs engaging with data collection clerks. Their rich interactions often resulted in software adaption, data management refinement, process flow improvements and a range of learning opportunities for all involved. Throughout these interactions there was always a common attractor, vision or ideology that was shared by CAS participants: the improvement of health services through the collection and use of information.

The ability of a system to integrate its information will grow as that system incorporates statistical-regularities within its environment (Tononi, 2004; Waser, 2011) and the majority of statistical regularities developed out of the need for improvement. When data at macros levels could no longer be effectively integrated, the system agents would make use of workshops or meetings to address their improvement needs. The need for direction was supported by a variety of committees and organising bodies (e.g. NHISSA), development partners, government agencies and even international funders. Their relationships, feedback loops and interactions with DOH staff all influenced the direction of system content, quality and utilization. Eventually these statistical regularities were adopted as standard practice and were written as standard operating procedures (SOPs). When a new executive leader in the DOH recognized the value of these underlying processes and data producing structures formal recognition was given by establishing of a 'plan and control' document of guidelines realized as the DHMIS policy of 2011. This policy formally protects the working mechanisms and processes that produce, manage and ensure integrity of data in the health system of South Africa.

5.3 Emergence of Self-Service BI

Research material into the area of intelligence describes cognitive capabilities in which goal realization is possible with minimal effort and across a wide range of circumstances. These cognitive capabilities rely on information processing to achieve desired outcomes. Information processing refers to information production by complex elements within an integration structure while maintaining a high degree of informational-relationships (Waser, 2011; Tononi, 2004). In terms of BI focus areas these requirements, capabilities and mechanisms can be related back to various components of the CAS information system. Its dynamic and flexible data model provided informational structure and content for the flow of data according to a set of standards. Governance measures for adhering to DQ and MDM standards (e.g. timelines for data submission, accuracy measurement practices, master data integrity reviews according to a schedule, etc) had evolved over the years. Inputs and support regarding the development of data processes came from role players such as development partners, state agencies, DOH leadership and committees which were formally enacted by the DHMIS policy of 2011. It took more than a decade for these processes to emerge as coherent patterns of "order" for the CAS's national level. Each vertical level had opportunity to learn from its parent level's MDM and DQ processes (e.g. provincial administrations relied on shared knowledge and understanding of national level processes to develop and implement their own data collection and reporting specifications).

The lower the amount of information processing required to influence outcomes so that they include the goal, the higher the level of intelligence (Waser, 2011). In terms of BI information processing capabilities refer to a wide range of working mechanisms. Transformed data should be readily accessible and pre-arranged according to criteria that correspond to outcome options for selection by decision makers. This assumes that data being collected and transformed can be arranged to partially or wholly meet the option selection criteria needs of decision makers. For this process to work efficiently it requires rapid access to information that is already integrated, transformed and prepared in real time (i.e. a data warehouse/mart is loaded with aggregated and ETL transformed data with required outputs prearranged according to a prioritized set of needs). The distribution of spread marts created access but without the sophisticated information processing needs that are required to arrange outputs according to decision making priorities. This has had a slowing effect on high value information access. This aspect of the information cycle requires effort and time to convert spread data into informed reports. The existing information distribution processes made entire data repositories available in bulk without any prioritization or arrangement of outputs according to decision making priority.

Improving information access and reducing information processing time would be addressed through the development of a provincial intranet web reporting system that would allow users to produce a variety of information outputs arranged according to vertical, horizontal and health data dimensions. Because data mart distribution would take place each month according to a strict schedule ETL transformed data could be loaded into a centralized DBMS. The development of a web interface that integrated with this DBMS would provide 'predefined analytic perspectives' resulting in a first of its kind 'health intelligence' platform for the DOH. These 'predefined analytic perspectives' included stock exchange style information outputs (see Figure 26) that demonstrated month-on-month performance, pre-processed calculations for populations and health facility count ratios (see Annexure G) as well as GIS integration.

The platform would be used exclusively for the provision and presentation of output data. Through a series of interactions the national level DOH became aware of this platform and its underlying processes (i.e. the routine health information system known as DHIS). Data specification, collection, integration and transformation processes were already in place and were utilized. The NHIRD project began with secondment of data analysts and GIS specialists. The role of these specialists was to act as the BI team (see Figure 3) while developing new sources of data for the warehouse while the web platform was adopted as a self-service tool (see Figure 4). To date (2014) these data sources have grown to provide health insights along the lines of social and economic development, gender specific health concerns and regionalized profiles for the national health insurance initiative. The majority of these information products were designed to be portable and are not yet available in the self-service environment. What has emerged as a key area of interest is the GIS environment. Major investment in a commercial platform has seen significant work in this area. Integration of GIS within the reporting environment is still under developed. Since its adoption to the national level in 2011 the software and reporting capabilities of the web reporting platform have seen many new enhancements that were as a result of learnings from exchanges with local development partners, visiting specialists from the United Kingdom and local public health experts. New analytic perspectives are still being developed to reduce processing time for decision makers but those perspectives and decision making processes are currently under review.

5.4 Key Learnings

Based on experiences from this study a researcher wanting to utilize CAS theory to study IS phenomena requires access and time within their environment, a well-developed understanding of their IS subject domain and persistence. CAS theory requires us to engage with agents of a system in order to observe emerging patterns. Access to the environment is not guaranteed as agents are free to participate and choose with whom they engage. Generally the bigger the system under review the greater the number of agents available but this does not guarantee access or data quality. CAS theory may be well suited to action research projects. In this case study the environment was broad encompassing different levels of scale and its longitudinal nature extended back many years adding to the challenge of accessing information about its changing history. Numerous agents participated in the development of the underlying information processes and some remain present as active agents in the CAS today (2014). They persisted, acting as guides and problem solvers, more so than others. Identifying those core agents of the CAS should be a priority to researchers as it will help reduce research time and efforts.

Complex adaptive systems require attractors (Schneider & Somers, 2006; Sturmberg, O'Halloran & Martin, 2012) such as shared problems (e.g. fragmentation of HIS), common goals (e.g. creating integrated solutions for HIS) and vision (e.g. improving health services through the collection and use of information) around which interactions and information exchanges can occur naturally. Structure in CAS can be partially or temporarily replaced by attractors which may eventually lead to the creation of naturally forming structures and processes. Systems take time to develop but they do so under the correct conditions, e.g. freedom to participate and interact with other agents, freedom to develop relationships and trust with agents of other systems, freedom to share information through interactions, and freedom to learn sometimes from mistakes.

A CAS requires correct leadership and guidance to ensure progress and learning. Consider Figure 29 from the point of view that complex systems require constant nudges (from the top) to keep agents within the zone of complexity. This requires managers who understand CAS thinking. Systems cannot be allowed to stagnate as this deprives agents from learning and adapting their internal rules. When systems undergo rapid changes (e.g. agents with accumulated history and experience in the CAS leave or get replaced) information sharing is important to ensure smooth continuation as was learned by the provincial information collapse example. BI process development took many generations to get right. Master data (set) management processes have seen numerous expansion and contraction phases over the years. Many of the learnings from one sub-system have been shared as incremental learnings for the whole. Data quality initiatives and approaches are still being perfected as there are many persistent challenges with routine health information systems. Well-designed data warehouses and marts have provided great stability to the case study CAS. Data collection and integration processes have been difficult to manage across the numerous egional implementations but security measures related to the locking of master data sets (as a side effect of an information system audit) has had a stabilizing effect on data quality.

Based on this case study Routine Health Information Systems are an essential component of any functional HIS or HMIS. They can exist as an operationalised data collection, integration and analysis architecture across different levels of a country's administrative hierarchy. In this case study the RHIS of South Africa was embodied by the District Health Information System which started off as a small pilot project and quickly became an attractor for other health programmes. Its success was dependent on the shared vision of local agents who were able to utilize the DHIS as an instrument to expand that vision. Those agents that participated in its development supported the following BI processes:

- master data development and rationalization when appropriate (MDM);
- data collection, validation and integration (data collection and DQ);

- data transformation and distribution through portable spread marts (AN and DW);

These processes helped take a prototype software application from a single clinic in the Western Cape to the NHIRD platform representing a variety of data sets for an entire country. This provides evidence that bottom up approaches that work to build systems of information can spread when everyone has equal access to the system's data.

The stability and integrity of RHIS data depend on its perceived importance by health managers, decision makers, development partners and resource allocation authorities. Formal data quality improvement processes were implemented only after audit findings revealed security and process flow concerns (e.g. audits by Treasury highlighted discrepancies in some regions), which took place only after DHIS data was used to assist in the allocation of health budgets. While South Africa is a developing country it is richer than most and able to provide for its own health budget allocations. The same cannot be said for other countries in the African region that are dependent on funding from foreign donors. "In SA the government has enough money to fund the DOH but in other developing countries governments don't have enough money to fund their ministries." Our resource allocation authority (Treasury) was empowered to conduct audits of processes and information systems. The results of these findings have led to vast improvements with regards to data quality of the case study information system. What appears to be missing is a focus on improving data quality at lower levels.

RHIS master data development (of indicators and data elements) must be inclusive, transparent and participatory for all levels involved in the collection, integration and use of information. Getting consensus on data specifications across all vertical levels and health programmes is surely a massive undertaking but has been proven to work in the case study. Well-functioning processes lead to a stronger system. "Its policy distribution, it's the running of training programmes. It's coordinated by the provinces, and cascaded down to districts, sub-districts and facilities. It also includes giving feedback; feedback on data quality, data performance and things like that."

System break-downs should be understood according to data flow, information exchange (local system knowledge) and load capacity as small changes at higher (decision making) levels can have big impacts on the entire system. "We need to do much more work to empower and support facility level staff. I would try help district and provincial program managers to develop their information use processes. Give them guidance on how to understand the data, determine if it's good quality, if the numbers make sense. Work needs to be done to change the attitude at facility level. The data quality can be changed by addressing this need.

The four BI focus areas highlighted by Tan et al. (2011) in Table 1 intertwine heavily throughout this case study. From the early beginnings of DHIS version 1.3 there was strong emphasis on master data (set) specifications and alignment to organisational hierarchy structures. Data collection processes have been supported by information quality (IQ) assessment tools have relied heavily on analytic capabilities. The aggregation and transformation of data into data marts made use of extremely generic and flexible analytic processes that helped produce standardized pivot table files which supported data distribution and information sharing across South Africa's DOH. The stabilization of these processes took more than fourteen years to materialize in the DHMIS policy which has formally recognized the need for structures based on processes that have evolved within the complex environment of South Africa's public health system (see Annexure I).

6. Conclusions & Implications

6.1 Conclusion

The NHIRD project unit is comprised of different people all working towards a common goal: to support problem-solving through the analysis of local information (stored and obtained from the National Data Warehouse). It houses different types of data, both qualitative & quantitative. The project (unit) consists of a self-service BI tool and a BI team comprised of problem solvers, GIS experts and master data analysts. The BI team's main tasks include the development of GIS capabilities and tools, the development of information products in response to BU requests which sometimes require the design of innovative analysis methodologies. The self-service BI tool contains custom developed visual analytic outputs along with standardized reports and graphs. These analytic tools are still evolving and seem to be a great area for future development. The self-service platform is considered a "corporate asset" in the National Department of Health and is treated and protected accordingly. Perhaps access to this tool is overly restrictive but appears to be out of concern for information exploitation or misuse. These processes are taking place within the public health system of a developing country in Africa and not a corporateenvironment.

Meadows (1999) stated that, in systems, leverage points are points of control or power that are often well known by many but are not always understood and sometimes pushed in the wrong direction. A major learning from this case study is that this statement is true and that people (who seek to increase their “fitness” in the environment) may unknowingly sabotage the evolution of systems by perceiving many of the available options as leverage points. This ‘in over your head’ symptom had resulted in the exhaustion of resources and energy because efforts were pushed in all directions instead of being more focused. It is impossible for one person to have knowledge of an entire complex adaptive system (Hammer et al., 2010) therefore relationships need to be nurtured and maintained to ensure maximum information flows (and learnings) between components of a CAS.

BI processes depend on supportive policies. Many of their supporting mechanisms are defined by policies and the failure of any one mechanism can be attributed to a variety of symptoms, e.g. disempowerment, lack of resources, lack of capacity, delays, etc. A disempowered individual who is delegated with implementing a BI-supportive process may unwillingly weaken the local and entire system. A lack of resources or capacity may result in poor quality data or delay access and outputs of information. The absence of supportive policies, empowered and motivated individuals, resources and capacity all have weakening effects on a BI system and delay benefits realisation.

Numerous support processes made the accumulation and integration of data in the National Data Warehouse possible. Data coordination, mandated data-requirements, process flows, the delegation of power, etc, all influence the inflow of data to the National Data Warehouse. The arrangements of these mechanisms have taken many years to reach the point where i) master data can be accessed and shared through a centralized reference point, ii) data quality is at a point where health service budgets allocations are being trusted because underlying data is of good quality, iii) aggregated routine and indicator data from each and every health facility in the country can be accessed through a SSBI interface in the National Data Warehouse, iv) analytics is being used to develop deeper understandings of health-indicator performance, and v) analytics is developing to the point where performance management through target and threshold setting is being planned.

Was it always this way in South Africa’s HMIS? In its early days the answer was a resounding ‘NO’. In the early 1990’s, in the absence of information flows and relationships as a result of historic policies, health system agents worked together with development partners to reach a point where master data specification, data collection and flow, and data distribution processes would stabilize. After numerous generations and iterations of learning an information management culture established itself within this system which led to the creation of self-regulating processes and structures. Soon after this resource allocation authorities began utilizing data from this system to allocate health-budgets.

Policies and self-regulating practices would emerge to support the coordination of master data development, data collection, data quality, data integration and data presentation processes. The environment changed. There would be no more parallel information silos working in competition to one another all competing for funding and serving different needs. Data would be integrated across health programmes forming macro level overviews. Decision making would be coordinated and resource allocations would become evidence based.

The DHIS has provided the DOH with a dynamic and flexible tool to create a learning path for other developing countries. This has taken place over many information management iterations and cycles. Management practices and decision making have changed because of various incremental leaps that appear to have focused primarily on the specification and management approach towards master data design. The strength of BI processes in public health has been anchored by many different participants of a CAS who are all attracted to the vision of a work-force empowered by information sharing and knowledge development.

Can BI processes from the corporate domain be relevant in low resource settings? Much of the findings have demonstrated the presence of MDM review processes, DQ improvement workshops, the distributed DW and spreadsheet phenomena known as ‘spread marts’ and the establishment of SSBI. Over the past fifteen to twenty years these processes have been supported by agents of the health system to make the flow of data up to a national level possible. There has been an increasing interest in performance management metrics (e.g. target setting) which is becoming an area of priority and increasing interest to executive level administrators in health. Structures, policies and dedicated staff have ensured the flow and sharing of information within this CAS.

6.2 Limitations, Recommendations and Areas for Further Research

This case study looked into routine health information systems in a developing country and considered RHIS from a public health system perspective. It excluded the evolution of private sector health information processes as the private sector is currently isolated from public sector information flows. BI processes were discussed from a corporate or private enterprise perspective which tends to follow a “best of current practice”. These best practices appear to be common knowledge in the private sector while the public sector persists with its own approach to technology and best practices adoption.

Private sector organizations are more capable of heavy investment in technology. The evolution of BI processes in public health has arrived through a different approach, one that appears to have placed public health workers at the centre of a RHIS evolutionary track. The private sector evolutionary track may be considerably different without emphasis on routine information. This may be an area of interest to public health or health system academics.

Routine Health Information Systems offer a wide variety of opportunities to low resource countries. They stand as integrated repositories for health data across programmes and administrative concerns. They provide master data management tools that can be linked to or accessed by other types of electronic health systems. They create information infrastructure that would otherwise exist in disconnected silos. They can act as the glue to create a unified health information system which may lead to the realization of BI associated benefits.

BI processes in health work well when health workers are included in process-designs. In order for process-design to work health workers require the presence of flexible tools and technologies that can be adapted to suite a variety of needs under a variety of circumstances. This flexibility can act as an attractor to information system stakeholders. Flexibility and complexity have complemented one another in the case study and complexity could even be seen as an attractor that helped people to share experiences, interact with others and develop relationships across regular boundaries of comfort.

There are many opportunities for studying BI processes in South Africa’s public health system that are still accessible as future areas of research. These include information product development processes at sub-national levels to assist provincial and sub-provincial decision makers. Further areas of interest must include the mapping of decision processes in public health to available information producing systems. This may require the development of knowledge management processes but will remain an area of interest for further investigation.

When you consider that a RHIS collects time series (TS) data it becomes apparent that more work should be done to develop new visual analytic methods to help facilitate the multiplier effect described by Andrienko and Andrienko (2013). TS analysis methods for visual analytics have been shown to be an area of interest to users of the case study SSBI tool.

Kimball and Ross (2011) discuss the data warehouse metadata framework describing its composition according to data and metadata. While enterprise information management approaches place a different emphasis on data components, e.g. master data and their associated sub-classes together with transactional or operational data, it must be considered that information architectures in developing countries are enhanced by the presence of routine health information systems. As integrators of information an RHIS may provide leverage in the architecture space. Special considerations need to be made for situations where country level information architectures evolve out of data warehouse implementations.

One final and noteworthy concept worth mentioning is the health site that produces real time data collected without the need for paper registers. This is a concept of the imagination to many public health specialists interviewed in this case study. It has been confirmed many times over in discussions with colleagues and in at least one interview. It will remain an area of interest for this researcher.

6.3 Concluding Remarks

It is unnatural for there to be a competitive nature between public health service providers so as to draw people to their services. In the private sector health services performance is stimulated by financial rewards but the reward mechanisms in public health are different. This raises interesting considerations around the purpose of BI-driven processes for health system management. Private enterprises often utilize information intelligence to identify market gaps or to seek competitive advantages but how would public health management best take advantage of “intelligence”?

Business strategies typically focus on growing market-share or becoming leaders in a particular field. In public health there are no similar ‘drivers’ because reward mechanisms are different. In terms of health services management, efforts to improve delivery and outcomes of health services seem to be driven by a desire to achieve optimal “efficiency” which raises questions such as: what are the real (information) attractors that drive health system managers? The case study SSBI tool is being positioned as a “business intelligence” platform being able to deliver actionable “intelligence” in the hopes that it will stimulate the development of an information culture. Having sophisticated analytic tools does not mean it will be used it unless there is an attractor. The question remains: what are those attractors? Could these tools be designed to allow provinces to obtain their maximum budget-allocation possible? Could they be used to understand health patterns better? Perhaps public and private sectors could both use their tools to match supply with demand to achieve an optimal allocation of resources.

7. References

AbouZahr, C., & Boerma, T. (2005). Health information systems: The foundations of public health. *Bulletin of the World Health Organization, 83*(8), 578-583.

Anderson, P. (1999). Complexity theory and organization science. *Organization Science, 10*(3), 216-232.

Andrienko, N., & Andrienko, G. (2013). A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery, 27*(1), 55-83.

Alkire, S. (2010). Human Development: Definitions, Critiques, and Related Concepts. Human Development Research Paper HDRP/2010/01. New York: United Nations Development Programme, Human Development Report Office.

Aqil, A., Lippeveld, T., & Hozumi, D. (2009). PRISM framework: A paradigm shift for designing, strengthening and evaluating routine health information systems. *Health Policy and Planning, 24*(3), 217-228.

Arnott, D., & Pervan, G. (2005). A critical analysis of decision support systems research. *Journal of Information Technology, 20*(2), 67-87.

Bai, X., Li, P., Li, H., & Song, X. (2010). Enterprise master data manage project practice. *International Conference on Computer Application and System Modeling (ICCASM 2010),* North University of China, Taiyuan, China. *, 10.* 220-223.

Baškarada, A., & Koronios, A. (2013). Data, information, knowledge, wisdom (DIKW): A semiotic theoretical and empirical exploration of the hierarchy and its quality dimension. *Australasian Journal of Information Systems, 18*(1), 5-24.

Berson, A., & Dubov, L. (2007). *Master data management and customer data integration for a global enterprise*. New York: McGraw-Hill.

Boisot, M., & Child, J. (1999). Organizations as adaptive systems in complex environments: The case of China. *Organization Science*, 10(3), 237-252.

Bose, R. (2008). Competitive intelligence process and tools for intelligence analysis. *Industrial Management & Data Systems, 108*(4), 510-528.

Braa, J., Hanseth, O., Heywood, A., Mohammed, W., & Shaw, V. (2007). Developing health information systems in developing countries: The flexible standards strategy. *MIS Quarterly, 31*(2), 381-402.

Bucher, T., Gericke, A., & Sigg, S. (2009). Process-centric business intelligence. *Business Process Management Journal, 15*(3), 408-429.

Campbell, M., Machin, D., & Walters, S. (2010). *Medical statistics: A textbook for the health sciences.* John Wiley & Sons.

Capra, F. (1996). *The web of life*. London: HarperCollins Publishers.

Chapter 6: Health Information White Paper for the Transformation of the Health System in South Africa. (2014). Retrieved June 14, 2014, from http://www.healthlink.org.za/pphc/Phila/chap06.htm

Chaudhuri, S., Dayal, U., & Narasayy, V. (2011). An overview of business intelligence technology. *Communications of the ACM, 54*(8), 88-98.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165-1188.

Cleven, A., & Wortmann, F. (2010). Uncovering four strategies to approach master data management. *43rd Hawaii International Conference on System Sciences (HICSS),* pp. 1-10.

Cohen, M. (1999). Commentary on the organization science special issue on complexity. *Organization Science, 10*(3), 373-376.

Cosma, S., Văleanu, M., Cosma, D., Vasilescu, D., & Moldovan, G. (2013). Efficient data organisation in distributed computer systems using data warehouse. *International Journal of Computers, Communications & Control, 8*(3), 366-374.

Cuzzocrea, A., Song, I., & Davis, K. (2011). Analytics over large-scale multidimensional data: The big data revolution! *Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP,* Glasgow, Scotland. pp. 101-104.

Daft, R. (1992). *Organization theory and design* (4th ed.). St. Paul: West Publishing.

Dayal, U., Castellanos, M., Simitsis, A., & Wilkinson, K. (2009, March). Data integration flows for business intelligence. In Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (pp. 1-11). ACM.

Dreibelbis, A., Hechler, E., Milman, I., Oberhofer, M., van Run, P., & Wolfson, D. (2008). *Enterprise master data management: An SOA approach to managing core information* Pearson Education.

Gosain, A., Nagpal, S., & Sabharwal, S. (2011). Quality metrics for conceptual models for data warehouse focusing on dimension hierarchies. *ACM SIGSOFT Software Engineering Notes,36*(4), 1-5.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*(1), 13-23.

Hammer, R. J., Edwards, J. S., & Tapinos, E. (2012). Examining the strategy development process through the lens of complex adaptive systems theory. *Journal of the Operational Research Society*, 63(7), 909-919.

Hart, M. (2009). Business intelligence projects in second year information systems courses. Proceedings of the 2009 Annual Conference of the Southern African Computer Lecturers' Association, Mpekweni Beach Resort, South Africa. pp. 68-75.

Haux, R. (2006). Health information systems–past, present, future. *International Journal of Medical Informatics, 75*(3), 268-281.

Hemmings, J., & Wilkinson, J. (2003). What is a public health observatory? *Journal of Epidemiology and Community Health, 57*(5), 324-326

Inmon, W. H., Strauss, D., & Neushloss, G. (2010). *DW 2.0: The architecture for the next generation of data warehousing: The architecture for the next generation of data warehousing.* Morgan Kaufmann.

Hyde, K. (2000). Recognising deductive processes in qualitative research. *Qualitative Market Research: An International Journal, 3*(2), 82-89.

IBM (2011). Global CIO study: The essential CIO. IBM Institute for Business Value.

Kahaner, L. (1998). *Competitive intelligence: How to gather, analyse and use information to move your business to the top*. New York: Touchstone.

Kappelman, L., McLean, E., Luftman, J., & Johnson, V. (2013). Key Issues of IT Organizations and Their Leadership: The 2013 SIM IT Trends Study, *MIS Quarterly Executive*, December 2013, 227-240.

Kimball, R., & Ross, m. (2011). *The data warehouse toolkit: The complete guide to dimensional modelling.* John Wiley & Sons.

Klein, H., & Myers, M. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS Quarterly,* 67-93.

Koronis, A., & Yeoh, W. (2010). Critical success factors for business intelligence systems. *Journal of Computer Information Systems, 50*(3), 23-32.

Leischow, S., & Milstein, B. (2006). Systems thinking and modeling for public health practice. *American Journal of Public Health, 96*(3), 403-405.

Lippeveld, T. (Ed.). (2000). *Design and implementation of health information systems*. France: World Health Organization.

Lippeveld, T. (2001). Routine health information systems: The glue of a unified health system. *Keynote Address at the Workshop on Issues and Innovation in Routine Health Information in Developing Countries, Potomac, March,* pp. 14-16.

Lucas, A. (2011). Corporate data quality management: Towards a meta-framework. *International Conference on Management and Service Science (MASS),*pp. 1-6.

Luftman, J., Zadeh, H., Derksen, B., Santana, M., Rigoni, E. & Huang, Z. (2013). Key information technology and management issues 2012–2013: an international study. *Journal of Information Technology* 28, 354-366.

March, S. T., & Hevner, A. R. (2007). Integrated decision support systems: A data warehousing perspective.*Decision Support Systems, 43*(3), 1031-1043.

Meadows, D. (1999). Leverage points: Places to Intervene in a System. The Sustainability Institute.

Merali, Y., Papadopoulos, T., & Nadkarni, T. (2012). Information systems strategy: Past, present, future?. *The Journal of Strategic Information Systems*, 21(2), 125-153.

Meso, P., & Jain, R. (2006). Agile software development: adaptive systems principles and best practices. *Information Systems Management*, 23(3), 19-30.

Michalewicz, Z., Schmidt, M., Michalewicz, M., & Chiriac, C. (2006). *Adaptive business intelligence* (pp. 37-46). Berlin Heidelberg: Springer.

Mikroyannidis, A., & Theodoulidis, B. (2010). Ontology management and evolution for business intelligence. *International Journal of Information Management*, 30(6), 559-566.

Mills, A., Rorty, M., & Werhane, P. (2003). Complexity and the role of ethics in health care. *Emergence*, 5(3), 6-21.

Nan, N. (2011). Capturing bottom-up information technology use processes: A complex adaptive systems model. *MIS Quarterly, 35*(2), 505-532.

National Department of Health. (2011). District Health Management Information System (DHMIS) Policy. South Africa. Retrieved June 9, 2014, from http://www.hst.org.za/sites/default/files/District%20Health%20Management%20Information%20System%20Policy.pdf

Negash, S. (2004). Business Intelligence. *Communications of the Association for Information Systems*, 13, 77–195.

Olszak, C., & Ewa, Z. (2006). Business intelligence systems in the holistic infrastructure development supporting decision-making in organisations. *Interdisciplinary Journal of Information, Knowledge, and Management, 1*, 47-58.

Olszak, C., & Ziemba, E. (2007). Approach to building and implementing business intelligence systems. *Interdisciplinary Journal of Information, Knowledge, and Management, 2*, 134-148.

Otto, B., & Reichert, A. (2010). Organizing master data management: Findings from an expert survey. *Proceedings of the 2010 ACM Symposium on Applied Computing,* pp. 106-110.

Ployhart, R., & Vandenberg, R. (2010). Longitudinal research: The theory, design, and analysis of change. *Journal of Management, 36*(1), 94-120.

Popovič, A., Turk, T., & Jaklič, J. (2010). Conceptual model of business value of business intelligence systems. Management: *Journal of Contemporary Management Issues*, 15(1), 5-30.

Ritchie, J., Lewis, J., Nicholls, C., & Ormston, R. (2013). *Qualitative research practice: A guide for social science students and researchers.* Sage.

Russel, S., Haddad, M., Bruni, M., & Granger, M. (2010). Organic evolution and the capability maturity of business intelligence. *AMCIS 2010 Proceedings,*(501)

Sallam, R., Richardson, J., Hagerty, J., & Hostmann, B. (2011). Magic quadrant for business intelligence platforms. Stamford, CT. Gartner Group.

Schneider, M., & Somers, M. (2006). Organizations as complex adaptive systems: Implications of complexity theory for leadership research. *The Leadership Quarterly, 17*(4), 351-365.

Shaw, V. (2009). A complex inspired approach to co-evolutionary hospital management information systems. Philosophiae Doctor (Ph.D.), Faculty of Mathematics and Natural Sciences, University of Oslo.

Smith, H., & McKeen, J. (2008). Developments in practice XXX: Master data management: Salvation or snake oil? *Communications of the Association for Information Systems, 23*(1), 4.

Stacey, R. (1996). *Complexity and creativity in organizations.* San Francisco, CA, US: Berrett-Koehler Publishers.

Stebbins, R. (2001). *Exploratory research in the social sciences*. Thousand Oaks, CA: Sage University Papers Series on Qualitative Research Methods.

Sturmberg, J., O'Halloran, D., & Martin, C. (2012). Understanding health system reform–a complex adaptive systems perspective. *Journal of Evaluation in Clinical Practice*, 18(1), 202-208.

Tan, C., Sim, Y., & Yeoh, W. (2011). A maturity model of enterprise business intelligence. *Communications of the IBIMA, 2011*(417812), 1-11.

Thomas, G. (2011). A typology for the case study in social science following a review of definition, discourse, and structure. *Qualitative Inquiry, 17*(6), 511-521.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5(1), 42.

Tozer, G. (1999). *Metadata management for information control and business success.* Artech House, Inc.

Trochim, W. M., Cabrera, D. A., Milstein, B., Gallagher, R. S., & Leischow, S. J. (2006). Practical challenges of systems thinking and modeling in public health. *American Journal of Public Health*, 96(3), 538.

Walsham, G. (2006). Doing interpretive research. *European Journal of Information Systems*, 15(3), 320-330.

Waser, M. (2011). Architectural requirements & implications of consciousness, self, and" free will". *Bica,* pp. 438-443.

Watson, H., & Wixom, B. (2007). The current state of business intelligence. *Computer, 40*(9), 96-99.

Weiskopf, N., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association, 20*(1), 144-151.

Yin, R. K. (2002). Case Study Research: Design and Methods, (Applied Social Research Methods, Vol. 5).

Yu, E., Lapouchnian, A., & Deng, S. (2013). Adapting to uncertain and evolving enterprise requirements. In Proc. 7th IEEE International Conference on Research Challenges in Information Science, 155-166.

Zeng, L., Li, L., & Duan, L. (2012). Business intelligence in enterprise computing environment. *Information Technology and Management, 13*(4), 297-310.

8. Annexures

Annexure A. Sample Data elements collected in the Routine Health Information System (DHIS)

Data Element	Definition	Comment
Antenatal 1st visit before 20 weeks	A first visit by a pregnant woman to a health facility that occurs before 20 weeks after conception	The actual protocol followed during the visit might vary but it should include: Relevant screening procedures, laboratory tests (e.g. for syphilis), counselling and health promotion (often done in groups)
Born alive before arrival at facility	Live born baby to a woman who had intended/booked a facility delivery but delivered before arrival and reached a health facility within 72 hours for normal post-delivery care (BBAs)	Multiple births are counted as several live births
Inpatient beds - total	All inpatient beds that are approved for use within the health facility	These beds are the usual accommodation where the patient will spend most of his/her stay. The patient may temporarily leave this unit of accommodation for surgery or examinations or treatment, but the bed will be kept for them and they will return to it after the treatment or examination. This temporary accommodation is not counted as a bed
Maternal death in facility	A maternal death in facility is the death of a woman while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and the site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management, but not from accidental or incidental causes - that occur while in a health facility	This should be collected in all units of a health facility
Patient Day Equivalent	Weighted data element as proxy for estimating resources for all types of patients in terms of inpatient days	([Emergency headcount total] x 33% + [OPD headcount - total] x 33%) + ([Day patients - total] x 50%) + [Inpatient days - total
PHC headcount 5 years and older	All individual clients five years (60 months) and older seen for Primary Health Care	If a delivery occurs at a CHC or MOU that does not admit clients as inpatients, ONLY count the mother as a PHC headcount and count the baby as a headcount ONLY if immunisation services are provided to the baby. If the client is recorded as a headcount at the registration point, BUT no observations of the client is done and the client then leaves the facility without receiving the service that he/she is supposed to receive, a headcount should NOT be recorded. In cases where a PHC service is rendered on an individual basis to a client at another place than a PHC facility by a team consisting of professional practitioners and HCBC workers, all data for the relevant data elements should be recorded by the professional practitioners and not by the HCBC worker/s. If medication for a client is collected by a Home Community Based Care (HCBC) Worker or family member/friend, a headcount of the client whose medication is collected CANNOT be registered as a headcount at the registration point

Annexure B. Sample Health Indicators defined as part of the National Indicator Data Set

Indicator	Definition	Numerator	Denominator
Cervical cancer screening coverage	Cervical smears in women 30 years and older as a proportion of 10% of the female population 30 years and older	Cervical cancer screening in woman 30 years and older	Population 30 years and older female / 10
Child under 5 years diarrhoea case fatality rate	Proportion of children under 5 years admitted with diarrhoea who died	Child under 5 years with diarrhoea death	Child under 5 years with diarrhoea admitted
Child under 5 years diarrhoea with dehydration incidence	Children under 5 years newly diagnosed with diarrhoea with dehydration per 1,000 children under 5 years in the population	Child under 5 years diarrhoea with dehydration new	Population under 5 years
Cost per patient-day rate	Average cost per patient day equivalent	Total Expenditure	Number of Patient Days
Delivery by caesarean section rate	Delivery by caesarean section as proportion of total deliveries in health facilities	Delivery by caesarean section	Delivery in facility total
Immunisation coverage under 1 year	Proportion children under 1 year who completed their primary course of immunisation	Immunised fully under 1 year new	Population under 1 year
Inpatient bed utilization rate – total	Inpatient bed days used as proportion of maximum Inpatient bed days available. (Number of Inpatient beds X days in period)	Inpatient days + Half day clients	Inpatient beds - Total
Inpatient neonatal death rate	Proportion of children 28 days admitted/separated who died during their stay in the facility as a proportion of Live birth in facility	Inpatient death neonatal	Live birth in facility
Maternal mortality in facility ratio	Women who died in hospital as a result of childbearing, during pregnancy or within 42 days of delivery or termination of pregnancy, per 100,000 live births in facility	Maternal death in facility	Live birth in facility
PHC utilisation rate	Average number of PHC visits per person per year in the population	PHC headcount total	Population total
TB MDR death rate	Proportion MDR-TB patients who died during treatment period	TB MDR client death during treatment	TB MDR confirmed client initiated on treatment

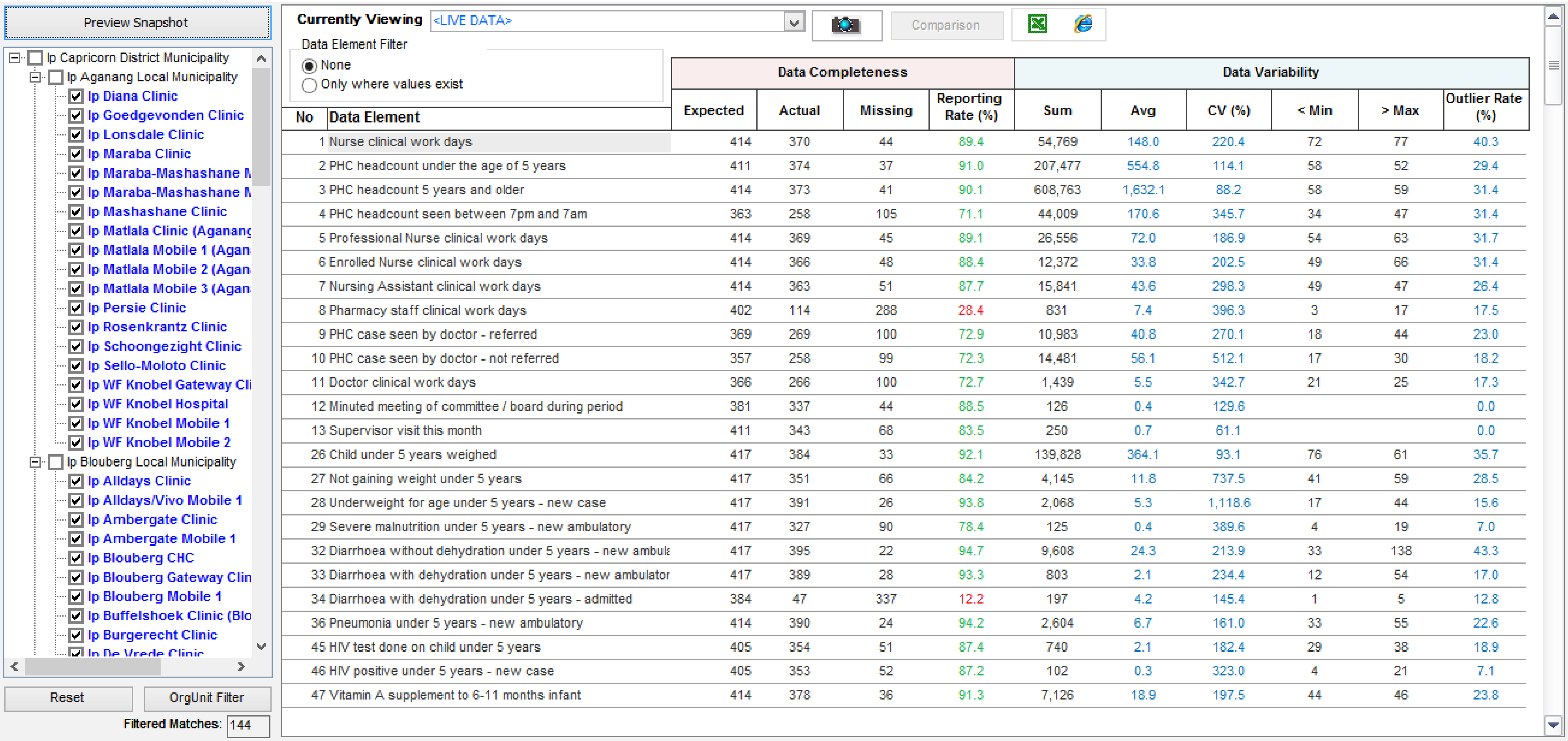
Annexure C. Vertical Dimension modelled across different Horizontal Data Sets making up the DHIS system

Vertical Dimension	Data Set	National Integrated Data Set	Ward Based Outreach Team Data Set	Antiretroviral Therapy Data Set	Environmental Health Services Data Set	Integrated School Health Programme Data Set
1	count descriptor	**1** National	**1** National	**1** National	**1** National	**1** National
2	count descriptor	**9** Province	**9** Province	**9** Province	**9** Province	**9** Province
3	count descriptor	**52** District	**52** District	**52** District	**52** District	**52** District
4	count descriptor	**247** Sub-District	**247** Sub-District	**247** Sub-District	**247** Sub-District	**247** Sub-District
5	count descriptor	**8340** Facility	**8340** Facility	**8340** Facility	EHS Service	School
6	count descriptor	Reporting Unit	Political Ward			
7	count descriptor		Outreach Team			

Annexure D. Different health service connected to their regional administrative offices

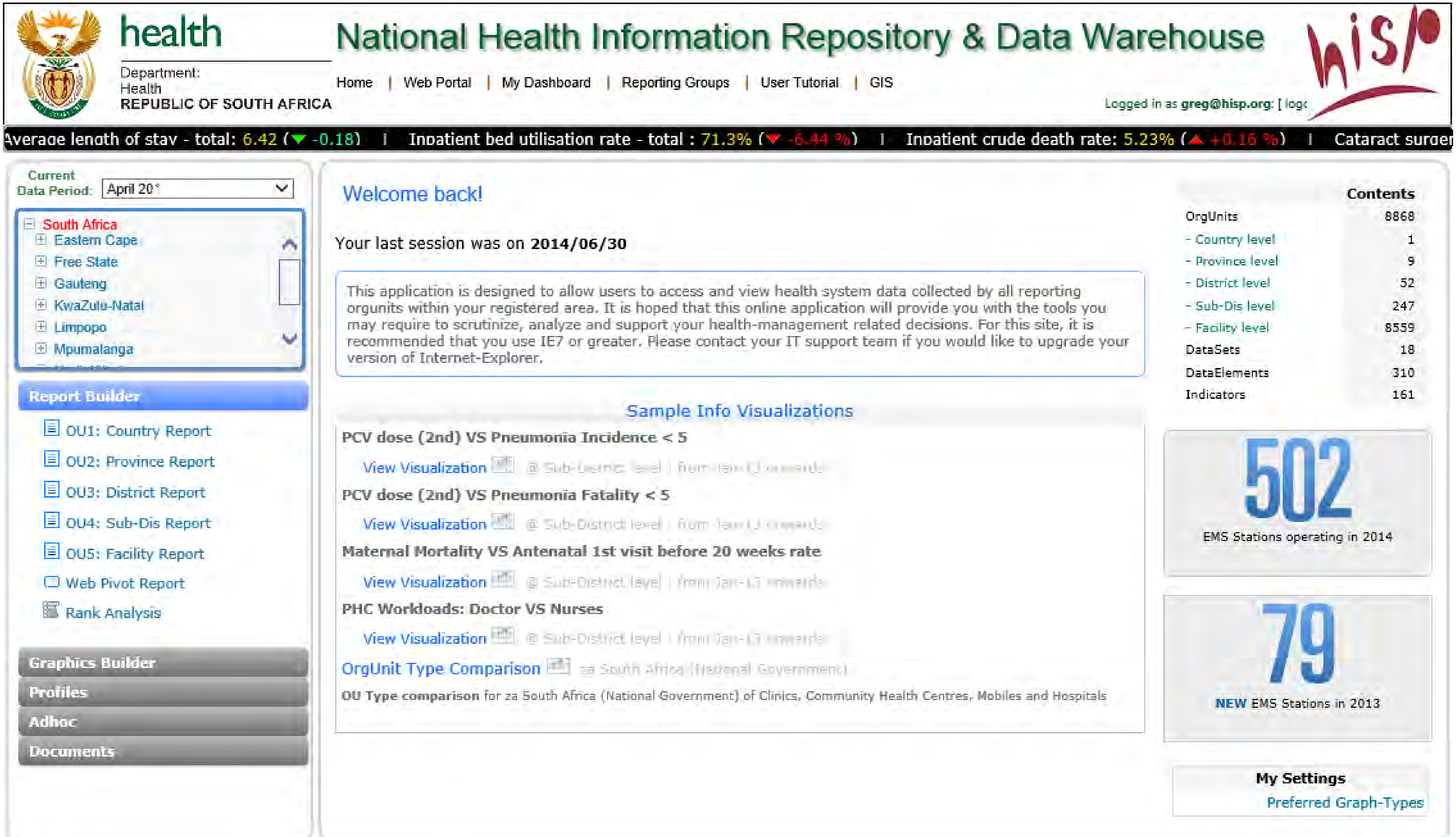
Special Note: The KZN provincial office is shown connected to its district-level offices and to the National Department of Health in Tshwane

Annexure E. Screenshot of DQ Completeness Tool

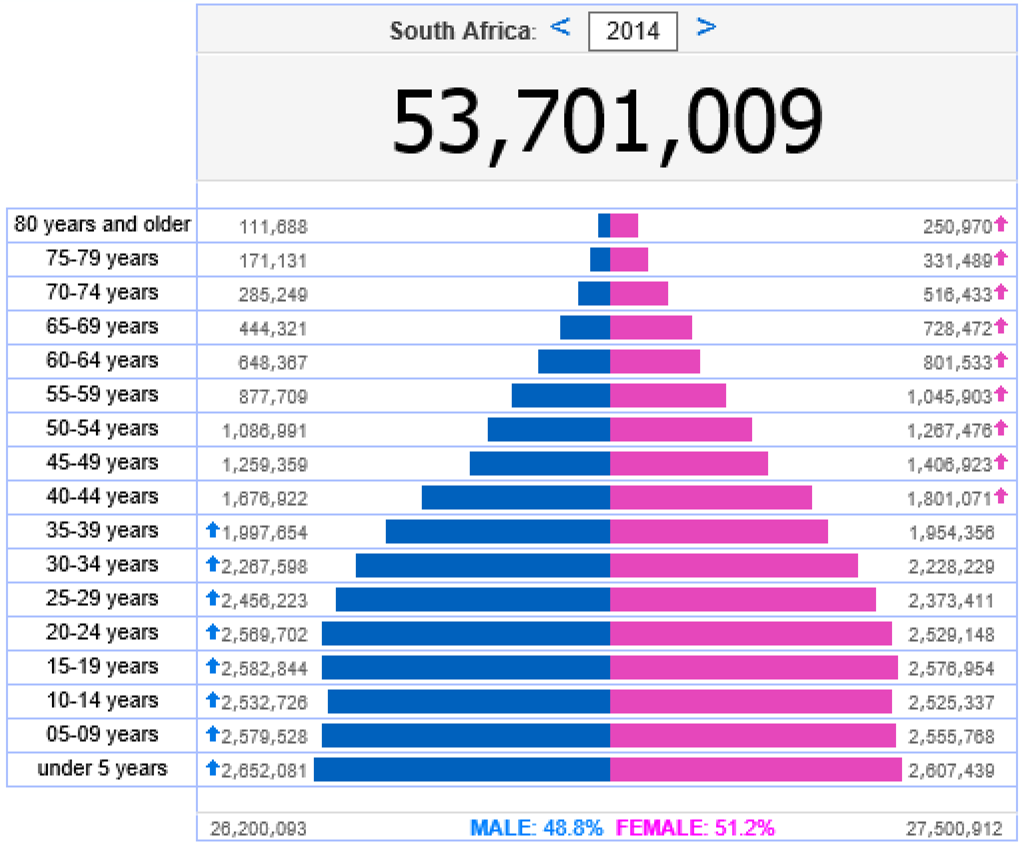


No	Data Element	Data Completeness				Data Variability					
		Expected	Actual	Missing	Reporting Rate (%)	Sum	Avg	CV (%)	< Min	> Max	Outlier Rate (%)
1	Nurse clinical work days	414	370	44	89.4	54,769	148.0	220.4	72	77	40.3
2	PHC headcount under the age of 5 years	411	374	37	91.0	207,477	554.8	114.1	58	52	29.4
3	PHC headcount 5 years and older	414	373	41	90.1	608,763	1,632.1	88.2	58	59	31.4
4	PHC headcount seen between 7pm and 7am	363	258	105	71.1	44,009	170.6	345.7	34	47	31.4
5	Professional Nurse clinical work days	414	369	45	89.1	26,556	72.0	186.9	54	63	31.7
6	Enrolled Nurse clinical work days	414	366	48	88.4	12,372	33.8	202.5	49	66	31.4
7	Nursing Assistant clinical work days	414	363	51	87.7	15,841	43.6	298.3	49	47	26.4
8	Pharmacy staff clinical work days	402	114	288	28.4	831	7.4	396.3	3	17	17.5
9	PHC case seen by doctor - referred	369	269	100	72.9	10,983	40.8	270.1	18	44	23.0
10	PHC case seen by doctor - not referred	357	258	99	72.3	14,481	56.1	512.1	17	30	18.2
11	Doctor clinical work days	366	266	100	72.7	1,439	5.5	342.7	21	25	17.3
12	Minuted meeting of committee / board during period	381	337	44	88.5	126	0.4	129.6			0.0
13	Supervisor visit this month	411	343	68	83.5	250	0.7	61.1			0.0
26	Child under 5 years weighed	417	384	33	92.1	139,828	364.1	93.1	76	61	35.7
27	Not gaining weight under 5 years	417	351	66	84.2	4,145	11.8	737.5	41	59	28.5
28	Underweight for age under 5 years - new case	417	391	26	93.8	2,068	5.3	1,118.6	17	44	15.6
29	Severe malnutrition under 5 years - new ambulatory	417	327	90	78.4	125	0.4	389.6	4	19	7.0
32	Diarrhoea without dehydration under 5 years - new ambul	417	395	22	94.7	9,608	24.3	213.9	33	138	43.3
33	Diarrhoea with dehydration under 5 years - new ambulator	417	389	28	93.3	803	2.1	234.4	12	54	17.0
34	Diarrhoea with dehydration under 5 years - admitted	384	47	337	12.2	197	4.2	145.4	1	5	12.8
36	Pneumonia under 5 years - new ambulatory	414	390	24	94.2	2,604	6.7	161.0	33	55	22.6
45	HIV test done on child under 5 years	405	354	51	87.4	740	2.1	182.4	29	38	18.9
46	HIV positive under 5 years - new case	405	353	52	87.2	102	0.3	323.0	4	21	7.1
47	Vitamin A supplement to 6-11 months infant	414	378	36	91.3	7,126	18.9	197.5	44	46	23.8

Annexure F. Screenshot of Self-Service BI Interface



Annexure G. Screenshot of NHIRD population pyramid with organisational unit ratios per population



za South Africa (National Government)	Number of Facilities	2014 estimated Population per Healthcare Unit
PHC	4507	11,915 :1
EMS	502	106,974 :1
Private PHC	259	207,340 :1
District Hospitals	255	210,592 :1
Private Hospitals	145	370,352 :1
Specialised Hospital	76	706,592 :1
Regional Hospitals	48	1,118,771 :1
Tertiary / National Hospital	27	1,988,926 :1

Annexure H. Sample Questions used to Guide Interviews

Q1. How was the NHIRD project established and what were its major influences?

Q2. Who are the most important decision makers when it comes to public health data?

Q3. What processes support the collection of data used by the NHIRD?

Q4. What are the major areas of concern for public health right now?

Q5. Who manages the data specification and development process of the NIDS?

Q6. How did the DHIS make the change from a prototype to becoming a national information system that gets used by Treasury?

Q7. In the context of a developing country, with limited infrastructure, what would you say is the core data set that any health minister or official should have access to at all times?

Q8. How are data workshops coordinated and who decides when it’s time to have one?

Q9. In your view, who are the most important decision makers with regards to public health in a developing country?

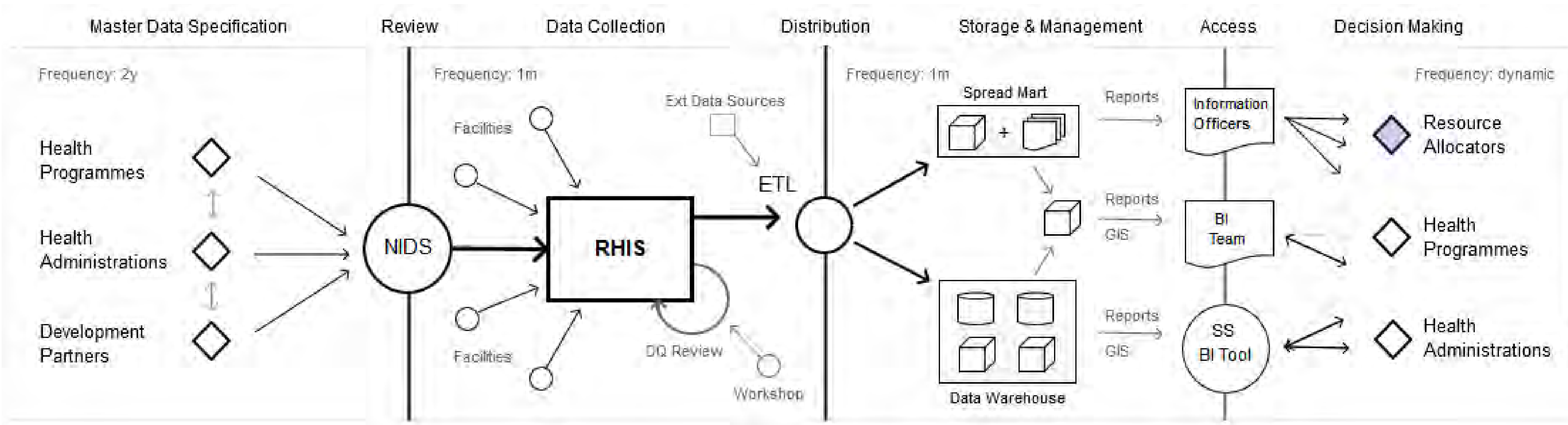
Q10. The DHIS data collection and integration process takes place once per month. How has data quality been implemented to ensure integrity, accuracy and timeliness of the data?

Q11. Please explain why the national data dictionary prototype developed in 2003 did not get adopted at a national level.

Q12. How did Treasury get involved with DHIS data?

Q13. How is performance management conducted in health?

Annexure I. DHMIS Arrived Process Flow with Emerging SS BI Tool



Special note: this has been adapted from Cosma, Văleanu, Cosma, Vasilescu & Moldovan, (2013), see Figure 12.